# 计算机辅助下的肺结核诊断

刘云，吴宇寰，张世辰，刘丽，吴敏，程明明

**摘要**—结核病（Tuberculosis, TB）是全球主要的健康威胁之一，每年导致数百万人死亡。尽管早期诊断和治疗可以极大地提高存活的机会，但在发展中国家仍然是一个重大挑战。最近，使用深度学习的计算机辅助结核病诊断（Computer-aided Tuberculosis Diagnosis, CTD）显示出了希望，但受限于有限的训练数据，进展受到了阻碍。为了解决这个问题，我们建立了一个大规模数据集，名为结核病 X 光片（Tuberculosis X-ray, TBX11K）数据集，其中包含 11,200 张胸部 X 光片（Chest X-ray, CXR）图像，以及相应的用于结核病区域的边界框标注。该数据集使得能够训练出用于高质量 CTD 的复杂检测器。此外，我们提出了一个强大的基准模型，SymFormer，用于同时进行 CXR 图像分类和结核感染区域的检测。利用 X 光片图像的双侧对称性属性（Bilateral symmetry property），SymFormer 开发了对称搜索注意力（Symmetric Search Attention, SymAttention）以学习有区别的特征。由于 CXR 图像可能不严格遵循双侧对称性属性，我们还提出了对称位置编码（Symmetric Positional Encoding, SPE）来通过特征重校准促进 SymAttention。为了促进对 CTD 的未来研究，我们通过引入评估指标、评估从现有检测器改进的基准模型，并进行在线挑战来建立了一个基准。实验证明，SymFormer 在 TBX11K 数据集上取得了最先进的性能。数据、代码和模型将在https://github.com/yun-liu/Tuberculosis上发布。

**关键词**—肺结核，肺结核诊断，肺结核检测，对称搜索注意力，对称位置编码

## 1 引言

结核病（Tuberculosis, TB）是一种普遍传染病，几个世纪以来一直是死亡和患病的第二主要原因，通常在艾滋病毒（HIV）之后 [3], [4]。尽管 2020 年全球爆发了新冠疫情，结核病仍然每年影响 1000 万人，并导致 140 万人死亡 [5]，使其成为继新冠病毒后的第二致命传染病。结核病主要影响呼吸系统，由结核分枝杆菌引起，通过打喷嚏、剧烈咳嗽或其他传播感染细菌的手段传播。因此，结核病通常通过呼吸道在肺部发生。免疫受损的个体，包括 HIV 感染者和发展中国家的营养不良者，使这一问题更加严重。

在没有适当治疗的情况下，结核病患者的死亡率仍然极高。然而，结核病的早期诊断可以通过相应的抗生素治疗显著提高康复率 [6]–[8]。由于结核病传播迅速，早期诊断在控制感染传播方面也起着至关重要的作用 [7]。多药耐药结核病的出现强调了及时准确的诊断方法对监测临床治疗进展的紧迫需要 [9]。然而，结核病的诊断仍然是一个重大挑战 [6]–[8], [10]–[13]。具体而言，用于结核病诊断的金标准涉及对痰样本和细菌培养的显微镜检查，以鉴定结核分枝杆菌 [12], [13]。为了确保检查过程的安全性，培养结核分枝杆菌需要生物安全级别 3（Biosafety Level-3, BSL-3）实验室。这个过程通常需要几个月 [6], [12], [13]。加剧问题的是，许多发展中国家和资源匮乏社区的医院缺乏建立 *BSL-3* 设施的必要基础设施。

另一方面，X 射线成像是目前医学图像检查中最常见和数据密集的筛查方法。胸部 X 光片（Chest X-ray, CXR）可以迅速检测由肺结核引起的肺部异常，使其成为结核病筛查的广泛使用工具。世界卫生组织还推荐将 CXR 作为结核病筛查的初始步骤 [14]。通过 CXR 的早期诊断显著有助于早期结核病的检测、治疗和预防疾病传播 [6], [11], [14]–[16]。然而，即使经验丰富的放射科医生在 CXR 图像中可能无法识别结核病感染，因为人眼对 CXR 图像中的许多细节的敏感性有限。我们的人体研究显示，与黄金标准相比，来自顶级医院的经验丰富的放射科医生的准确率仅为 *68.7%*。

由于深度学习具有卓越的表示学习能力，它在各个领域已经超过了人类，如人脸识别 [17]、图像分类 [18]、目标检测 [19], [20]、边缘检测 [21], [22] 和医学图像分析 [23]–[25]。有理由期望将深度学习强大的潜力应用于使用 CXR 进行结核病诊断。深度学习可以自动定位精确的结核病感染部位，每天 24 小时不知疲倦地工作，不像人类。然而，深度学习依赖于大量的训练数据，而现有的结核病数据集无法提供，如表1所示。由于由于高昂的成本和隐私考虑，收集大规模的结核病 CXR 数据具有挑战性，现有的结核病数据集只有几百张 CXR 图像。公开可用的 CXR 数据稀缺阻碍了深度学习在改进计算机辅助结核病诊断（Computer-aided Tuberculosis Diagnosis, CTD）性能方面的成功应用。

为了在全球范围内部署 CTD 系统以帮助结核病患者，


- 本文为 *[1]* 的翻译版。
- 该工作的大部分内容是当刘云在南开大学 *VCIP* 实验室时所做，他现在的单位是新加坡科技研究局（*A*STAR*）、信息通信研究所（*I2R*）。*(邮箱: vagrantlyun@gmail.com)*
- 吴宇寰来自新加坡科技研究局（*A*STAR*）、高性能计算研究所（*IHPC*）。*(邮箱: wu_yuhuan@ihpc.a-star.edu.sg)*
- 张世辰和程明明来自南开大学 *VCIP* 实验室。*(邮箱: zhangshichen@mail.nankai.edu.cn 和 cmm@nankai.edu.cn)*
- 刘丽来自于国防科技大学、电子信息与技术学院。*(邮箱: liuli_nudt@nudt.edu.cn)*
- 吴敏来自新加坡科技研究局（*A*STAR*）、信息通信研究所（*I2R*）。*(邮箱: wumin@i2r.a-star.edu.sg)*
- 通信作者: 刘丽. *(邮箱: liuli_nudt@nudt.edu.cn)*

- 这项工作的初步版本已在 *CVPR* 上发表（口头报告）*[2]*。

表 1
**公开可用结核病数据集概要。我们的数据集大小约为先前最大数据集的 17 倍。此外，我们的数据集使用边界框标注结核感染区域，而不仅仅是图像级别的标签。**

| 数据集 | 发表年份 | 类别数 | 标注 | 样本数 |
|---|---|---|---|---|
| MC [26] | 2014 | 2 | 图像级 | 138 |
| Shenzhen [26] | 2014 | 2 | 图像级 | 662 |
| DA [7] | 2014 | 2 | 图像级 | 156 |
| DB [7] | 2014 | 2 | 图像级 | 150 |
| TBX11K (我们的) | - | 4 | 边界框 | 11,200 |

首先需要解决数据不足的问题。在本文中，我们通过与重点医院的长期合作，向社区贡献了一个大规模的**结核病 X 光片（Tuberculosis X-ray, TBX11K）**数据集。这个新的 TBX11K 数据集在几个方面超过了以前的 CTD 数据集：i）与以前的公共数据集 [7], [26] 只包含数十或数百张 CXR 图像不同，TBX11K 包含 11,200 张 CXR 图像，大约是现有最大数据集，即深圳数据集 [26] 的 17 倍，使得训练深度网络成为可能；ii）与以前数据集中的图像级别标注不同，TBX11K 使用边界框标注结核感染区域，使未来的 CTD 方法能够识别结核病表现并检测结核区域，协助放射科医生做出明确的诊断；iii）TBX11K 包括四个类别：健康、患病但非结核病、活动性结核病和陈旧性结核病，与以前数据集中的二元分类（即结核病或非结核病）不同，使未来的 CTD 系统能够适应更复杂的现实场景，并为人们提供更详细的疾病分析。TBX11K 数据集中的每个 CXR 图像都经过结核病诊断的金标准（即诊断微生物学）测试，并由来自主要医院的经验丰富的放射科医生进行标注。TBX11K 数据集已由数据提供者去标识，并获得相关机构的豁免，使其能够公开共享以促进未来 CTD 研究。

基于我们的 TBX11K 数据集，我们提出了一个简单而有效的 CTD 框架，称为 **SymFormer**。受 CXR 图像中固有的双侧对称性属性（Bilateral symmetry property）启发，SymFormer 利用这一属性来增强对 CXR 图像的表征。双侧对称性属性表示胸部左右两侧的相似或相同外观，表明一种对称模式。这一属性在改善 CXR 图像解释方面非常有价值。例如，如果胸部的一侧存在肿块或实变而另一侧没有，可能表示该区域存在问题。为了解决这一属性，SymFormer 引入了新颖的**对称搜索注意力（Symmetric Search Attention, SymAttention）**，用于从 CXR 图像中学习有区别的特征。由于 CXR 图像可能不严格遵循双侧对称性，我们还提出了**对称位置编码（Symmetric Positional Encoding, SPE）**，通过特征校准促进 SymAttention。SymFormer 通过在结核病感染区域检测器上添加分类头并使用两阶段训练图表，同时进行 CXR 图像分类和结核病感染区域检测。

为了促进未来 CTD 研究，我们在 TBX11K 数据集上建立了一个基准。具体而言，我们将图像分类和目标检测的评估指标调整为 CTD，以规范 CTD 的评估。我们还使用 TBX11K 的测试数据启动了一个在线挑战，通过保持测试数据的真实值私有，使未来对 CTD 的比较更加公平。此外，我们通过改进现有流行的目标检测器构建了几个强大的 CTD 基线模型。广泛的比较证明了 SymFormer 相对于这些基线模型的优越性。

与初步的会议版本相比 [2]，我们通过为 CTD 提出了一个新颖的 SymFormer 框架并通过大量实验证明其有效性，进行了丰富的扩展。总体而言，本文的贡献有三个方面：

- 我们建立了一个大规模的 CTD 数据集，TBX11K，比以前的结核病数据集更大、标注更好、更真实，使得能够训练深度神经网络，同时进行多类 CXR 图像分类和结核病感染区域检测，而不仅仅是以前结核病数据集中的二元 CXR 分类。
- 我们提出了一个简单而有效的 CTD 框架，即 SymFormer，包括新颖的**对称搜索注意力（SymAttention）**和**对称位置编码（SPE）**，以利用 CXR 图像的双侧对称性属性，显著提高了 CTD 相对于基线模型的性能。
- 我们在 TBX11K 数据集上建立了一个 CTD 基准，引入了评估指标，评估了几个从现有目标检测器改进的基线，并启动了一个在线挑战，预计为未来的研究打下一个良好的基础。

## 2 相关工作

在本节中，我们首先回顾了以前的结核病数据集，然后对现有的 CTD 研究进行了回顾。由于我们提出的 CTD 方法 SymFormer 使用视觉 Transformer 的自注意力机制，我们还讨论了医学图像中视觉 Transformer 的最新进展。

### 2.1 结核病数据集

由于结核病数据非常私密，使用金标准诊断结核病很困难，公开可用的结核病数据集非常有限。我们在表1中提供了公开可用结核病数据集的概要。Jaeger 等人 [26] 建立了两个用于结核病诊断的 CXR 数据集。蒙哥马利县胸部 X 光片数据集（Montgomery County chest X-ray set, MC）[26] 通过与美国马里兰州蒙哥马利县卫生和人类服务部合作收集。MC 数据集包含 138 张 CXR 图像，其中 80 张为健康病例，58 张为表现出结核病症状的病例。深圳胸部 X 光片数据集（Shenzhen chest X-ray set, Shenzhen）[26] 通过与深圳市第三人民医院、广东医学院合作收集。深圳数据集由 326 个正常病例和 336 个表现出结核病症状的病例组成，总共 662 张 CXR 图像。Chauhan 等人 [7] 建立了两个数据集，即 DA 和 DB，这两个数据集分别来自新德里国立结核病与呼吸系统疾病研究所的两台不同 X 光片机器。DA 由训练集（52 张结核病和 52 张非结核病 CXR 图像）和独立测试集（26 张结核病和 26 张非结核病 CXR 图像）组成。DB 包含 100 张训练 CXR 图像（50

张结核病和 50 张非结核病）和 50 张测试 CXR 图像（25 张结核病和 25 张非结核病）。需要注意的是，这四个数据集都只包含用于二元 CXR 图像分类的图像级别标注。

这些数据集过小，无法训练深度神经网络，因此尽管深度学习在计算机视觉领域取得了许多成功故事，但最近关于 CTD 的研究一直受到限制。另一方面，现有数据集仅具有图像级别的标注，因此我们无法使用先前的数据训练结核病检测器。为了帮助放射科医生做出准确的诊断，我们希望检测结核病感染区域，而不仅仅是图像级别的分类。因此，缺乏结核病数据阻碍了深度学习在实际 CTD 系统中取得成功的可能性，这些系统有望每年挽救数百万结核病患者。在本文中，我们建立了一个带有边界框标注的大规模数据集，用于训练深度神经网络，同时进行 CXR 图像分类和结核病感染区域检测。这个新数据集的提出有望有益于未来的 CTD 研究，并推动更实用的 CTD 系统的发展。

### 2.2 计算机辅助结核病诊断

由于数据不足，传统的 CTD 方法无法训练深度神经网络。因此，传统方法主要使用手工设计的特征，并为 CXR 图像分类训练二元分类器。Jaeger 等人 [6] 首先使用图割分割方法 [27] 分割肺部区域。然后，他们从肺部区域提取手工设计的纹理和形状特征。最后，他们应用一个二元分类器，即支持向量机（Support Vector Machine, SVM），将 CXR 图像分类为正常或异常。Candemir 等人 [11] 采用基于图像检索的患者特定自适应肺模型与非刚性配准驱动的强健肺分割方法，这对于传统肺特征提取是有帮助的 [6]。Chauhan 等人 [7] 实现了一个 MATLAB 工具箱 TB-Xpredict，该工具箱采用 Gist [28] 和 PHOG [29] 特征，用于区分结核病和非结核病的 CXR 图像，而不需要分割 [30], [31]。Karargyris 等人 [32] 提取形状特征描述肺的整体几何特征和纹理特征来表示图像特征。

与使用手工设计的特征不同，Lopes 等人 [10] 采用在 ImageNet [33] 上预训练的冻结卷积神经网络作为 CXR 图像的特征提取器。然后，他们训练 SVM 来分类提取的深度特征。Hwang 等人 [8] 使用私有数据集训练了一个 AlexNet [34] 进行二元分类（结核病和非结核病）。与这些私有数据集一样，[35] 中也使用了其他私有数据集进行图像分类网络的训练。然而，我们提出的大规模数据集 TBX11K 已经公开共享，以促进该领域的研究。有了我们的新数据集，我们提出了基于 Transformer 的 CTD 方法 SymFormer，用于同时进行 CXR 图像分类和结核病感染区域检测，它作为未来 CTD 研究的一个强有力的基线，实现了最先进的性能。

### 2.3 医学图像中的视觉 Transformer

Transformer [36] 最初是在自然语言处理（Natural Language Processing, NLP）中引入的，它具有很好的捕获长距离依赖关系的能力。将 Transformer 调整为视觉任务的先驱工作，如 ViT [37]、DeiT [38] 和 P2T [39]，表明 Transformer 网络可以在性能上超越广泛使用的卷积神经网络。因此，包括医学图像在内的计算机视觉领域对视觉 Transformer 的关注逐渐增加。各种工作已经致力于将视觉 Transformer 应用于医学图像分割 [40]–[44] 和医学图像分类 [45]–[49]。然而，相对于分割和分类，基于 Transformer 的技术在医学图像检测方面的采用相对滞后。

大多数利用视觉 Transformer 进行医学图像检测的研究主要建立在 DETR（Detection Transformer）框架 [50] 上。该领域的开创性工作是 COTR [51]，它包括用于特征提取的卷积神经网络、用于特征编码的混合 Convolution-Transformer 层、用于对象查询的 Transformer 解码层和用于息肉检测的前馈网络。Mathai 等人 [52] 采用 DETR [50] 来检测 T2 MRI 扫描中的淋巴结，可用于评估淋巴增生性疾病。Li 等人 [53] 提出了 Slice Attention Transformer（SATr）块，用于对不同计算机断层摄影（Computed Tomography, CT）切片之间的长距离依赖关系进行建模，该块可以插入基于卷积的模型以进行通用病变检测。有关医学图像中视觉 Transformer 更全面的回顾，请参考最近的综述文章 [54]–[56]。在本文中，我们提出了 SymFormer 用于使用 CXR 图像进行 CTD。SymFormer 同时进行 CXR 图像分类和结核病感染区域检测。它利用 SymAttention 来处理 CXR 图像的双侧对称性属性，并通过 SPE 进一步提升了性能。SymFormer 在 SymAttention 和 SPE 的支持下，表现出比最近流行的目标检测器基线更好的性能，表明其在 CTD 方面的优越性。

## 3 TBX11K 数据集

深度神经网络对大量训练数据有很高的依赖性，而现有的公开结核病数据集规模并不大，如表1所示。为解决这一问题，我们建立了一个名为 TBX11K 的综合大规模数据集，使得能够为 CTD 训练深度网络成为可能。在本节中，我们首先在 §3.1中描述了我们如何收集和标注 CXR 数据。接下来，在 §3.2中，我们展示了由有经验的放射科医生进行的人体研究的结果。最后，在 §3.3中，我们讨论了可以利用我们的 TBX11K 数据集探讨的潜在研究主题。

### 3.1 数据收集和标注

为了收集和标注数据，我们遵循四个主要步骤：i）建立分类法，ii）收集 CXR 数据，iii）专业数据标注，iv）数据集划分。我们将在下面详细介绍每个步骤。

#### *3.1.1* 建立分类法

当前的结核病数据集仅包括两个类别：结核病和非结核病，其中非结核病指的是健康病例。然而，在实践中，结核、肺不张、心脏肥大、积液、浸润、肿块、结节等的 CXR 图像的异常具有相似的模糊和不规则病灶等异常模式，这与几乎没有异常模式的健康 CXR 图像有很大的不同。因此，仅依赖健康的 CXR 作为负类别会导致在模型对于有很多患有非结核疾

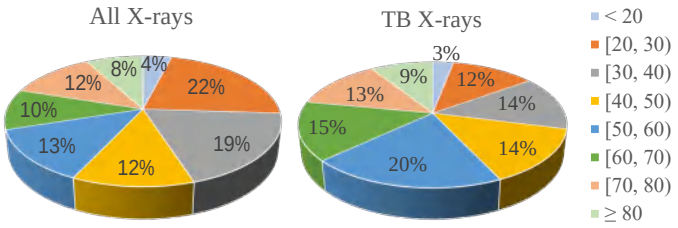


图 1. TBX11K 数据集的所有 X 光片的年龄分布以及结核 X 光片的年龄分布。

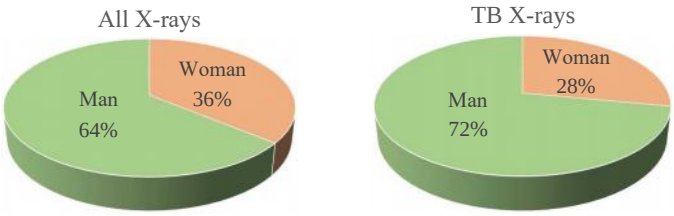


图 2. TBX11K 数据集的所有 X 光片的性别分布以及结核 X 光片的性别分布。

病的患者的临床场景的预测中产生大量假阳性。为了解决这个问题，促进 CTD 的实际应用，我们在我们的数据集中提出了一个新的类别，即患病但非结核病的类别。此外，区分活动性结核病和陈旧性结核病对于为患者提供适当的治疗至关重要。活动性结核病是由结核分枝杆菌感染或陈旧性结核病重新激活引起的，而陈旧性结核病患者既不患病也不具有传染性。因此，在我们的数据集中，我们将结核病分为两个类别：活动性结核病和陈旧性结核病。基于上述分析，提出的 TBX11K 数据集包括四个类别：健康、患病但非结核病、活动性结核病和陈旧性结核病。

### 3.1.2 数据收集

采集结核病 CXR 数据面临两个主要挑战：i）CXR 数据的隐私性较高，尤其是结核病 CXR 数据，几乎不可能让个人在不违法的情况下访问原始数据；ii）尽管全球范围内有数百万的结核病患者，但因结核分枝杆菌的复杂和漫长的检查过程，确定性地测试结核病 CXR 图像的数量仍然有限 [12], [13]。为解决这些挑战，我们与中国的顶级医院合作，收集了 CXR 数据。我们的 TBX11K 数据集包括 11,200 张 CXR 图像，包括 5,000 张健康病例，5,000 张患有非结核病病例和 1,200 张结核病病例。每张 CXR 图像对应一个独特的个体。在 1,200 张结核病 CXR 图像中，有 924 例活动性结核病病例，212 例陈旧性结核病病例，54 例同时具有活动性结核病和陈旧性结核病，以及 10 例无法当前医学条件下识别其结核类型的不确定病例。我们包括 5,000 例患有非结核病病例，以涵盖在临床场景中可能出现的广泛放射学疾病。数据提供者已对数据进行了去标识化处理，相关政府机构已免除了数据集，使其在法律上能够公开使用。

所有 CXR 图像的分辨率约为 3000 × 3000。每张 CXR

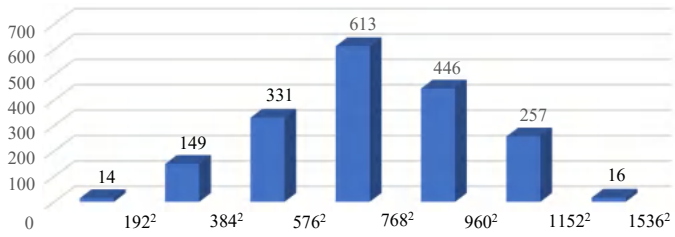


图 3. **TBX11K 数据集中结核边界框尺寸的分布**。每个柱代表边界框面积的特定范围。每个柱的左右值对应其面积范围，柱的高度表示该范围内的结核边界框的数量。应注意，TBX11K 数据集的 CXR 图像分辨率约为 3000 × 3000。

图像都附带有相应的年龄和性别信息，为结核病的诊断提供了全面的临床线索。年龄分布和性别分布分别显示在图1和图2中。从图1中可以看出，60% 的结核病患者年龄在 20 到 60 岁的范围内。只有很小比例的年轻个体（年龄小于 20 岁）感染结核病，具体为 3% 的结核病患者。这一发现与最近的医学研究一致 [57]–[59]。图2说明了大多数结核病患者是男性，与全球范围内结核病在男性中较女性更为普遍的临床观察一致 [60], [61]。我们使用边界框进行结核病感染区域的标注（在§3.1.3中介绍），结核边界框的大小分布显示在图3中。从图中可以看出，结核病感染区域的大小呈现出多样性，展示了不同程度的结核病病情严重程度。结合先前对分类法、类别分布、年龄分布、性别分布和结核病感染区域大小分布的分析，我们得出结论，新的 TBX11K 数据集既代表了一般人口，又符合现实临床场景。

### 3.1.3 专业的数据标注

我们的数据集包括经过金标准严格测试的 CXR 图像，提供了图像级别的标签。然而，虽然这种方法使我们能够将 CXR 图像归类为结核病与否，但它并未显示 CXR 图像中结核的具体位置或程度。检测这些结核病感染区域的能力对于使放射科医生能够做出明智的决策至关重要。目前，仅依赖图像级别的预测使人眼难以识别结核病感染区域，正如在临床检查中放射科医生的低准确性所证明的那样（参见 §3.2）。通过同时提供图像分类和结核定位结果，CTD 系统有望提高放射科医生做出明智决策的准确性和效率。

为了实现我们的目标，我们的 TBX11K 数据集包括了 CXR 图像中结核病感染区域的边界框标注。据我们所知，这是第一个专为结核病感染区域检测而设计的数据集。这些标注由来自顶级医院的有经验的放射科医生进行。具体来说，数据集中的每张结核病 CXR 图像首先由一位有 5-10 年结核病诊断经验的放射科医生标记。随后，另一位有超过 10 年结核病诊断经验的放射科医生审查了边界框标注。这些放射科医生不仅为结核病区域标注边界框，还为每个框识别结核病的类型（活动性或陈旧性）。为确保一致性，标记的结核病类型与金标准产生的图像级别标签进行了双重检查。如果存在不匹配，将 CXR 图像放置在未标记数据中进行重新标注，标注

表 2
**TBX11K 数据集的划分。**“活动性 & 陈旧性结核”指的是 CXR 图像中既有活动性结核又有陈旧性结核的情况；“活动性结核”指的是 CXR 图像中仅有活动性结核的情况；“陈旧性结核”指的是 CXR 图像中仅有陈旧性结核的情况；“不确定结核”指的是 CXR 图像中当前医学条件下无法识别结核感染类型的情况。

| | 类别 | Train | Val | Test | 合计 |
|---|---|---|---|---|---|
| 非结核 | 健康 | 3,000 | 800 | 1,200 | 5,000 |
| | 患病 & 非结核 | 3,000 | 800 | 1,200 | 5,000 |
| 结核 | 活动性结核 | 473 | 157 | 294 | 924 |
| | 陈旧性结核 | 104 | 36 | 72 | 212 |
| | 活动性 & 陈旧性结核 | 23 | 7 | 24 | 54 |
| | 不确定结核 | 0 | 0 | 10 | 10 |
| 合计 | | 6,600 | 1,800 | 2,800 | 11,200 |

者不知道之前哪张 CXR 图像被错误标记。如果一张 CXR 图像被错误标记两次，我们会告知标注者该 CXR 图像的金标准，并要求他们讨论如何重新标注。这个双重检查的过程确保了标注的边界框对于结核病感染区域的检测非常可靠。此外，非结核病 CXR 图像只有图像级别标签，没有边界框标注。TBX11K 数据集的示例显示在图6中，结核边界框大小的分布显示在图3中，表明大多数结核边界框在 $(384^2, 960^2]$ 的范围内。

#### *3.1.4* 数据集划分

我们将数据划分为三个子集：训练集、验证集和测试集，遵循表2中详细说明的划分方案。训练集和验证集的真实标签已公开，而测试集的真实标签保持机密。这是因为我们在我们的网站上使用测试数据启动了一项在线挑战。为了使数据集更具代表性，我们考虑了四种不同的结核病情况：i）仅包含活动性结核的 CXR 图像，ii）仅包含陈旧性结核的 CXR 图像，iii）同时包含活动性和陈旧性结核的 CXR 图像，以及 iv）在当前医学条件下无法识别结核类型的 CXR 图像。对于每种结核病的情况，我们在训练集、验证集和测试集中保持了 3 : 1 : 2 的结核病 CXR 图像数量比例。值得注意的是，不确定的结核病 CXR 图像已分配到测试集，使研究人员能够使用这 10 个不确定的 CXR 图像评估无关类别的结核检测。我们建议研究人员在训练集上训练模型，在验证集上调整超参数，并在使用训练集和验证集的合集进行重新训练后，在测试集上报告模型的性能。这种方法遵循科学实验设置，预计将产生可靠的结果。

### 3.2 放射科医生的人体研究

涉及放射科医生的人体研究是了解 CTD 在临床环境中发挥作用的关键组成部分。我们首先从新的 TBX11K 数据集的测试集中随机选择了 400 张 CXR 图像，其中包括 140 张健康的 CXR 图像，140 张患有非结核疾病的 CXR 图像，以及 120 张患有结核病的 CXR 图像。在这 120 张结核病的 CXR 图像中，有 63 张显示活动性结核，41 张显示陈旧性结核，15 张同时显示活动性和陈旧性结核，还有 1 张显示不确定的结核。接下来，我们邀请了一位在一家重点医院工作并具有超过 10 年工作经验的经验丰富的放射科医生，根据四个图像级别的类别（健康、患病但非结核病、活动性结核病和陈旧性结核病）为这些 CXR 图像进行标注。如果一张 CXR 图像同时显示活动性和陈旧性结核病的表现，医生会同时标注这两个标签。值得注意的是，这位放射科医生与标注原始数据集的医生不同。

与金标准相比，这位放射科医生的准确率仅为 68.7%。如果忽略活动性和陈旧性结核之间的区别，准确率提高到 84.8%，但在有效的临床治疗中区分结核的类型至关重要。这样的低准确率突显了结核病诊断、治疗和预防中的一项主要挑战。与自然彩色图像不同，CXR 图像是灰度的，通常具有模糊和不清晰的模式，使准确识别变得困难。不幸的是，在 BSL-3 实验室中使用金标准诊断结核病可能需要数月的时间，这在世界许多地方是不可行的。结核病诊断的挑战使其成为全球第二常见的传染病，仅次于 HIV。然而，我们将在即将进行的研究中展示，基于提出的 TBX11K 数据集训练的基于深度学习的 CTD 模型甚至可以显著优于有经验的放射科医生，为改进结核病诊断和治疗带来希望。

### 3.3 潜在的研究主题

展望未来，我们讨论一些与使用我们新开发的 TBX11K 数据集相关的潜在研究主题。

**同时进行分类和检测.** 我们的 TBX11K 数据集为进行 CTD 的研究打开了新的可能性，包括 CXR 图像分类和结核病感染区域检测。我们的测试集包含丰富的健康和非结核患病数据，使得能够模拟用于评估 CTD 系统的临床数据分布。我们认为，同时进行 CXR 图像分类和结核病感染区域检测系统的开发将是一个具有挑战性和迷人的研究主题，具有在结核病诊断中协助放射科医生的潜在应用。部署这样的系统最终可能提高结核病诊断和治疗的准确性和效率。

**不平衡的数据分布.** 除了同时进行检测和图像分类的挑战外，我们的 TBX11K 数据集还展示了不同类别之间的数据不平衡分布。然而，我们认为这种数据不平衡反映了实际的临床场景。当患者接受胸部检查时，他们可能正在经历不适或生病，增加了患病的可能性，而我们的数据集通过仅有 44.6% 的健康参与者捕捉到了这一现实。结核病仅仅是许多可能的胸部疾病之一，而我们的数据集通过仅有 10.7% 的参与者感染结核病，而 44.6% 的参与者患有非结核疾病，反映了这一现实。陈旧性结核病可能是由两种情况引起的：暴露于活动性结核病而被感染和在治疗后从活动性结核病转化而来。大多数陈旧性结核病病例是由于接触活动性结核病引起的。然而，患有陈旧性结核病的个体既不患病也不具有传染性，主动前去就医的可能性较低，这导致我们的数据集中活动性结核

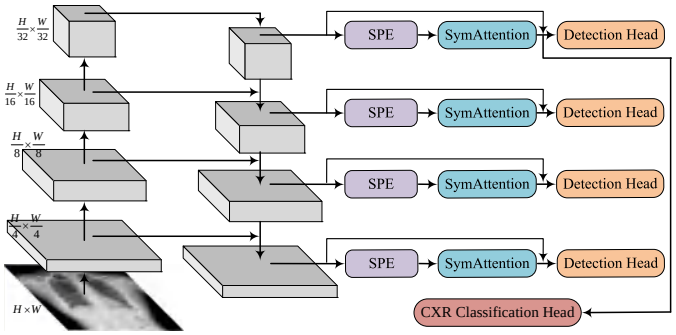


**图 4. 所提出的 SymFormer 框架示意图。**我们使用 FPN [62] 来生成特征金字塔。

病病例的数量比陈旧性结核病病例多。这种数据不平衡对未来的 CTD 方法提出了挑战，需要设计方法来解决实际问题。例如，需要开发在不平衡的 TBX11K 训练集上训练模型的方法，以提高结核病诊断的准确性。

**使用私有数据进行增量学习.** 增量学习是一种机器学习技术，涉及使用新数据更新模型的参数，而无需从头开始重新训练模型。鉴于围绕结核病 CXR 数据存在的高隐私问题，研究人员可能拥有无法公开的私有数据。在这种情况下，使用在 TBX11K 数据集上预训练的模型作为基础模型可能是有益的。研究人员可以利用增量学习，通过使用私有数据对预训练模型进行微调，从而提高模型对结核病诊断的准确性。因此，探讨使用新开发的 TBX11K 数据集进行 CTD 的增量学习的潜在研究方向也将是至关重要的。

## 4 我们的 SymFormer 框架

在这一部分，我们首先在 §4.1中概述我们的 SymFormer 框架。然后，我们在 §4.2中描述我们的**对称异常搜索 (Symmetric Abnormity Search, SAS)** 方法。SAS 包括两个组件：**对称位置编码 (Symmetric Positional Encoding, SPE)** (§4.2.1) 和**对称搜索注意力 (Symmetric Search Attention, SymAttention)** (§4.2.2)。接下来，我们介绍 SymFormer 的结核病诊断头 (§4.3)。最后，在 §4.4中，我们呈现了用于同时进行 CXR 图像分类和结核病感染区域检测的两阶段训练方式。

### 4.1 概述

我们在图4中展示了 SymFormer 的整体流程。SymFormer 包括三个部分：特征提取、对称异常搜索和结核病诊断头。我们将逐一详细介绍每个部分。

**特征提取.** 为了方便起见，我们以 ResNets [63] 作为特征提取的例子，因为它是社区公认的通用骨干网络。给定一张 CXR 图像作为输入，骨干网络在四个阶段输出特征，相对于输入尺寸，这些特征分别缩小了 1/4、1/8、1/16 和 1/32。由于结核病感染区域的大小和形状变化范围很大，因此从骨干网络中捕获多尺度特征至关重要。为了实现这一点，我们应用特征金字塔网络 (Feature Pyramid Network, FPN) [62] 在骨干网络上，生成一个特征金字塔，即在不同尺度上的特征图。我们将特征金字塔表示为 $\mathbf{F} = \{\mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3, \mathbf{F}_4\}$，其中 $\mathbf{F}_i \in \mathbb{R}^{C \times \frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}}} (i \in \{1,2,3,4\})$，$C$ 是特征维度，$H$ 和 $W$ 是输入 CXR 图像的高度和宽度。特征金字塔在实现跨不同特征层级的结核病感染检测方面非常有效。

**对称异常搜索.** SAS 模块用于增强提取的特征金字塔 $\mathbf{F}$。为了实现这一点，在 FPN [62] 的每个侧输出之后，我们加入了一个 SAS 模块，用于处理特征金字塔 $\mathbf{F}$ 中的每个特征图 $\mathbf{F}_i$。增强的特征金字塔表示为 $\hat{\mathbf{F}} = \{\hat{\mathbf{F}}_1, \hat{\mathbf{F}}_2, \hat{\mathbf{F}}_3, \hat{\mathbf{F}}_4\}$，其中 $\hat{\mathbf{F}}_i \in \mathbb{R}^{C \times \frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}}} (i \in \{1,2,3,4\})$。各侧输出处的 SAS 模块共享相同的权重，以减少网络参数数量。根据双侧对称性属性，正常 CXR 图像中的双侧对称区域应该看起来相似或相同。SAS 模块利用这一观点，在特征图的每个位置搜索对称位置，以确定它是否正常。SAS 模块由三个组件组成：SPE、SymAttention 和一个前馈网络。虽然 CXR 图像可能不是严格对称的，但 SPE 被设计为对特征进行重新校准，从而有益于 SymAttention 进行基于对称搜索的特征增强。

**结核病诊断头.** 我们将两种类型的结核病诊断头连接到由 SAS 模块增强的特征金字塔 $\hat{\mathbf{F}}$ 上，用于进行结核病感染区域检测和 CXR 图像分类。特征金字塔 $\hat{\mathbf{F}}$ 中的每个特征图都被送入检测头，每个检测到的边界框预期覆盖一个结核病感染区域。然而，在结核病感染区域检测过程中，对于非结核病 CXR 图像引入误报的风险较大，这会导致放射科医生为了进行临床诊断而检查这些误报，造成不必要的成本。为了解决这个问题，我们将增强的特征金字塔的顶层特征图 $\hat{\mathbf{F}}_4$ 送入分类头，以确定 CXR 图像是否包含结核。如果 CXR 图像被分类为结核病，则放射科医生可以进一步检查检测到的结核病感染区域，进行更准确和详细的临床诊断。如果 CXR 图像被分类为非结核病，则无需放射科医生进一步检查检测到的区域。

### 4.2 对称异常搜索

双侧对称性是 CXR 图像的一个属性，即胸部左右两侧的结构看起来相似或相同。换句话说，如果在 CXR 图像的中心画一条线，线的两侧的结构应该是近似相同的大小和形状。由于这一属性在 CXR 图像的解释中起到关键作用，使放射科医生和临床医生能够识别肺部的不对称或异常。例如，肺的一侧存在肿块或实变而另一侧不存在，这可能表明该区域存在问题。然而，值得注意的是，在正常 CXR 图像中并不总是存在完美的双侧对称性，这取决于患者的姿势和相对于拍摄 CXR 图像时的 X 射线机的位置。

我们所提出的方法 SAS 利用了双侧对称性属性来增强 CXR 图像的特征表示。如上所述，CXR 图像中的肺可能不是严格对称的。为了解决这个问题，SAS 首先采用 SPE 进行特

征重新校准。然后，SymAttention 使用重新校准的特征图来搜索特征图中每个位置的对称相邻区域，其中对称相邻区域指的是给定位置的双侧对称位置的相邻区域。SymAttention 通过注意力以自适应的方式聚合对称相邻区域的特征。相邻区域的确定也是通过学习方式进行的。通过迫使每个位置查看对称相邻区域，正如双侧对称性属性所建议的那样，我们可以学到 CXR 图像用于 CTD 的判别特征。

#### 4.2.1 对称位置编码

为了在特征图的自注意计算中引入位置信息，我们必须将位置编码添加到特征图中。有两种类型的位置编码：绝对位置编码（Absolute Positional Encoding）和相对位置编码（Relative Positional Encoding）[36], [37]。我们的方法 SPE 基于绝对位置编码，我们的实验证明相对位置编码相对于我们的 SPE 来说效果较差，如 §6.3所示。广泛使用的绝对位置编码 [36], [37] 采用不同频率的正弦和余弦函数：

$$\begin{aligned}\mathbf{P}[pos, 2j] &= \sin(pos/10000^{\frac{2j}{C}}),\\ \mathbf{P}[pos, 2j+1] &= \cos(pos/10000^{\frac{2j}{C}}),\end{aligned} \tag{1}$$

其中 $pos$ 表示空间位置，$j$ 表示特征维度的索引。对于特征金字塔 $\mathbf{F}$ 中的每个输入特征图 $\mathbf{F}_i$，我们使用公式1计算相应的位置编码 $\mathbf{P}_i$。$\mathbf{P}_i$ 与 $\mathbf{F}_i$ 具有相同的形状，以便对它们进行求和。

正如前文所提到的，由于可能存在轻微的旋转和平移，CXR 图像可能不严格遵循双侧对称性属性。所提出的 SPE 通过特征重新校准来解决这个问题。SPE 首先将位置编码 $\mathbf{P}_i$ 分成两侧，即 $\mathbf{P}_i^{left}$ 和 $\mathbf{P}_i^{right}$，通过在 $\mathbf{P}_i$ 的中心画一条竖线。然后，我们使用空间变换网络（Spatial Transformer Networks, STN）[64] 和水平翻转将 $\mathbf{P}_i^{right}$ 转移到左侧。最后，我们沿着 $x$ 维度拼接变换后得到的左侧位置编码和 $\mathbf{P}_i^{right}$，形成输出 $\mathbf{P}_i^{sym}$。这个过程可以表示为：

$$\begin{aligned}\mathbf{T}_i &= \mathrm{STN}(\mathbf{F}_i; \Theta)\\ \mathbf{P}_i^{trans} &= \mathrm{Flip}_x(\mathcal{T}_\Theta(\mathbf{P}_i^{right}; \mathbf{T}_i)),\\ \mathbf{P}_i^{sym} &= \mathrm{Concat}_x(\mathbf{P}_i^{trans}, \mathbf{P}_i^{right}),\end{aligned} \tag{2}$$

其中 $\Theta$ 是 STN 的权重；$\mathbf{T}_i$ 是仿射变换矩阵；$\mathcal{T}_\Theta$ 表示仿射变换；$\mathrm{Flip}_x$ 代表水平翻转；$\mathrm{Concat}_x$ 表示沿着 $x$ 维度的拼接。在公式2中，通过交换 $\mathrm{Concat}_x$ 的输入顺序，$\mathbf{P}_i^{right}$ 可以被替换为 $\mathbf{P}_i^{left}$。然而，我们在 §6.3中的实验证明，相对于 $\mathbf{P}_i^{left}$，$\mathbf{P}_i^{right}$ 的性能略优。对于每个输入 $\mathbf{F}_i$（$i \in \{1, 2, 3, 4\}$），我们使用公式2计算相应的 $\mathbf{P}_i^{sym}$。使用 SPE $\mathbf{P}_i^{sym}$，我们通过以下方式重新校准输入特征图：

$$\mathbf{F}_i^{recalib} = \mathbf{F}_i + \mathbf{P}_i^{sym}. \tag{3}$$

输出 $\mathbf{F}_i^{recalib}$ 将有助于后续 SymAttention 的计算。

**STN 的微观设计.** 公式2中的空间变换是依赖于输入特征和位置编码的。我们将输入特征 $\mathbf{F}_i$ 输入到 STN 的小型网络中，以预测仿射变换矩阵 $\mathbf{T}_i$，然后用于仿射变换一侧的位置编码 $\mathbf{P}_i^{right}$。这个小型网络包括两个交替的最大池化和 Conv-ReLU 层。然后，在空间维度上执行一个扁平化操作，接着是一个多层感知机（Multilayer Perceptron, MLP）用于预测仿射矩阵。我们初始化 MLP 以确保初始仿射矩阵的仿射变换等效于恒等映射。

#### 4.2.2 对称搜索注意力

自注意力在各个领域中变得流行 [65]–[68]，因为它能够学习序列或图像中元素之间的关系 [36], [37]。在医学图像分析中，自注意力已经应用于识别图像中的相关特征并增强疾病检测。然而，传统的自注意力通过为每个参考位置计算注意力权重来进行全局关系建模，从而融合所有位置的特征。这种方法对于 CTD 与 CXR 图像可能不是最优的。具体而言，自然图像可以在各种情境中捕获，并包含各种对象和元素，因此全局关系建模有助于理解整个场景。然而，CXR 图像只在单一场景中描绘人体胸部，各个 CXR 图像之间的差异通常仅限于难以察觉的异常区域的存在。因此，全局关系建模对于 CXR 图像可能是多余的，限制了自注意力学习增强特征表示的相关关系的能力。这是因为神经网络很难自动识别成千上万个冗余位置中的少数几个相关位置。例如，在我们的实验中，我们观察到当用于区分 CTD 中难以区分的结核特征时，DETR 检测框架 [50] 无法收敛。

为了解决这一挑战，我们提出 SymAttention，利用双侧对称性属性帮助自注意力在 CXR 图像中识别相关位置。正如前文所述，放射科医生可以通过比较肺部两侧的双侧对称位置来诊断结核病。因此，CXR 图像中每个参考位置的相关位置是双侧对称位置。在此启发下，SymAttention 在左右肺部中搜索对称模式的特征，使得每个参考位置只关注参考位置的双侧对称位置周围的位置。给定特征图 $\mathbf{F}_i^{recalib}$，我们首先选择一小组关键采样位置，按照 Deformable DETR [69] 的方法进行。设 $K$ 表示选择的位置数量，$M$ 表示自注意力计算中的头数。选定位置的坐标偏移可以通过以下学习得到：

$$\Delta\mathbf{p}_i^x = \mathbf{W}_x^{pos}\mathbf{F}_i^{recalib}, \qquad \Delta\mathbf{p}_i^y = \mathbf{W}_y^{pos}\mathbf{F}_i^{recalib}, \tag{4}$$

其中 $\mathbf{W}_x^{pos}, \mathbf{W}_y^{pos} \in \mathbb{R}^{(M\times K)\times C}$ 是可训练的参数矩阵。注意力 $\mathbf{A}_i$ 和值 $\mathbf{F}_i^v$ 可以简单地通过以下计算得到：

$$\begin{aligned}\mathbf{A}_i &= \mathrm{Softmax}(\mathrm{Reshape}(\mathbf{W}^{att}\mathbf{F}_i^{recalib})),\\ \mathbf{F}_i^v &= \mathbf{W}^{value}\mathbf{F}_i^{recalib}\end{aligned} \tag{5}$$

其中 $\mathbf{W}^{att} \in \mathbb{R}^{(M\times K)\times C}$，$\mathbf{W}^{value} \in \mathbb{R}^{C\times C}$ 是可训练的参数矩阵，softmax 函数沿着 $K$ 的维度执行。然后，我们将 $\mathbf{F}_i^v$ 变形为：

$$\mathbf{F}_i^v \in \mathbb{R}^{C\times \frac{H}{2^{i+1}}\times \frac{W}{2^{i+1}}} \rightarrow \mathbf{F}_i^v \in \mathbb{R}^{M\times \frac{C}{M}\times \frac{H}{2^{i+1}}\times \frac{W}{2^{i+1}}}. \tag{6}$$

接下来，SymAttention 可以表示为：

$$\mathbf{F}_i^{att} = \text{Concat}_{m=1}^{M}(\sum_{k=1}^{K}(\mathbf{A}_i[m,k]\cdot\mathbf{F}_i^v[m,:,\mathbf{p}_i^y+\Delta\mathbf{p}_i^y[m,k], \underbrace{\frac{W}{2^{i+1}}-(\mathbf{p}_i^x+\Delta\mathbf{p}_i^x[m,k])+1}])), \tag{7}$$

其中 $\text{Concat}_{m=1}^{M}$ 表示对 $m$ 从 1 到 $M$ 的所有结果进行拼接。带有波浪线的公式项通过以垂直中心线为对称轴，将采样位置投影到双侧对称位置上，这是提出的 SymAttention 的核心。最后，为了便于优化，连接了一个残差连接，然后是一个 MLP：

$$\hat{\mathbf{F}}_i^{att} = \mathbf{W}^{proj}\mathbf{F}_i^{att}+\mathbf{F}_i, \qquad \hat{\mathbf{F}}_i = \text{MLP}(\hat{\mathbf{F}}_i^{att})+\hat{\mathbf{F}}_i^{att}, \tag{8}$$

其中 $\mathbf{W}^{proj}\in\mathbb{R}^{C\times C}$，$\hat{\mathbf{F}}_i$ 是如 §4.1中所述的增强后的输出。

在公式4 - 公式8中，每个参考位置都关注参考位置的双侧对称位置周围的一小组关键采样位置，而不仅仅是对称位置。关键采样位置是以学习的方式自动设置的。这确保了在比较左右肺的外观时具有良好的感受野。在我们的实验中，在没有任何约束的情况下，我们观察到学习得到的坐标偏移 $\Delta\mathbf{p}_i^x$ 通常在相应特征图宽度的 10% 范围内。因此，对于位于垂直中心线远处的点，它们将搜索对称侧的点进行特征聚合。对于接近垂直中心线的点，它们通常不对应肺部区域（见图6），这不会影响我们对于在肺部区域检测结核时进行对称搜索的假设。在本文中，我们经验性地设置 $M=8$ 和 $K=4$。假设 $N=\frac{H}{2^{i+1}}\times\frac{W}{2^{i+1}}$，计算复杂度可以表示为 $\mathcal{O}(NC^2)$。因此，SymAttention 对于应用于特征金字塔 $\mathbf{F}$ 非常高效灵活。

### 4.3 结核病诊断头

在 §4.1中，我们提到有两个结核病诊断头：结核病感染区域检测头和 CXR 图像分类头。在本节中，我们详细介绍它们。检测头基于 RetinaNet [70]，一个著名的单阶段目标检测器，由用于边界框分类和位置回归的两个分支组成。与自然图像中每个边界框覆盖一个对象的目标检测不同，我们系统中的每个边界框设计为覆盖一个结核病感染区域。检测头学习以两个类别进行结核病检测：活动性结核病和陈旧性结核病。在临床结核病筛查中，大多数 CXR 病例没有结核感染，这使得检测头容易引入假阳性。为了解决这个问题，我们添加了一个 CXR 图像分类头，用于同时进行 CXR 图像分类和结核病感染区域检测。如果将 CXR 图像分类为非结核病，则丢弃检测到的结核区域。为简单起见，我们堆叠了几个卷积层以及池化操作作为分类头。有五个顺序的卷积层，每个卷积层有 512 个输出通道和 ReLU 激活。然后，我们采用全局平均池化获取全局特征向量，然后是一个具有 3 个输出神经元的全连接层，用于分类为三个类别：健康、患病但非结核病、结核病。

### 4.4 两阶段训练方式

我们的 SymFormer 框架包括两个头，分别设计用于 CXR 图像分类和结核病感染区域检测。在临床设置中，非结核病病例的数量明显多于结核病病例。直接使用非结核病病例进行感染区域检测头的训练将导致纯背景监督过多。因此，同时训练分类和检测头并不是最佳选择。此外，CXR 图像仅描绘胸部的结构和器官，不像自然图像那样具有复杂和多样的背景。如果我们首先训练骨干网络和分类头，那么用于特征提取的骨干网络将过度拟合，并且不会很好地推广到感染区域检测。此外，图像分类主要关注全局特征，而感染区域检测需要用于结核病区域定位的细粒度特征。因此，首先训练图像分类也不是最佳选择。我们提出的方法在初始阶段仅使用结核病 CXR 图像训练骨干网络和感染区域检测头。然后，我们冻结骨干网络和检测头，使用所有的 CXR 图像来训练分类头。这种训练策略受益于检测标注提供的更具体的边界框标注，从而减轻了过拟合的风险。通过感染区域检测学到的细粒度特征也可以轻松地转移到 CXR 图像分类中。

## 5 实验设置

在本节中，我们首先在 §5.1中详细说明了所提出的 SymFormer 的实现细节。随后，在 §5.2中介绍了用于 CTD 的几个基线模型，并在 §5.3中讨论了用于 CTD 的评估指标。

### 5.1 实现细节

我们使用 PyTorch [71] 和开源的 `mmdetection` 框架 [72] 来实现 SymFormer。第一阶段的训练使用了 TBX11K `trainval`（`train` + `val`）集中的结核病 CXR 图像，而第二阶段的训练不仅使用了所有 TBX11K `trainval` CXR 图像，还使用了 MC [26] 和 Shenzhen [26] 数据集的随机一半，以及 DA [7] 和 DB [7] 数据集的训练集。MC [26] 和 Shenzhen [26] 数据集的另一半，以及 TBX11K，DA [7] 和 DB [7] 数据集的测试集用于评估 CXR 图像分类的性能。我们将 FPN 特征通道的数量（即 $C$）设为 256，与 RetinaNet [70] 一致。其他设置也遵循 RetinaNet。对于第一阶段的训练，我们使用的 batch size 为 8，对于基于 Deformable DETR 的模型 [69]，训练 50 个 epochs，对于其他模型，训练 24 个 epochs。对于第二阶段的训练，我们使用的 batch size 为 8，对所有模型进行 12 个 epochs 的训练。我们对于基于 Deformable DETR 的模型使用 AdamW 优化器，对于其他模型使用 SGD 优化器。为了增强数据，我们使用随机翻转。我们将用于训练和测试的 CXR 图像的大小调整为 $512\times512$。所有实验都是在 2 个 TITAN XP GPU 上进行的。更多详细信息请参阅我们的代码。

### 5.2 基线模型

如 §4.1中所讨论的，使用图像分类头可以显著减少临床结核病筛查中检测的假阳性。然而，现有的目标检测器不考虑背景图像，通常忽略没有边界框对象的图像 [70], [73]–[76]。直接使用这些检测器进行 CTD 会导致由于临床实践中非结核病 CXR 图像的数量庞大而产生大量假阳性。为了解决这个问

题，我们引入了一个分类头，以实现同时进行 CXR 图像分类和结核病感染区域检测，其中 CXR 图像分类的结果用于过滤掉检测的假阳性。

为了实现这一点，我们重构几个著名的目标检测器，包括 SSD [73]、RetinaNet [70]、Faster R-CNN [75]、FCOS [74] 和 Deformable DETR [69]，以实现同时进行 CXR 图像分类和结核病感染区域检测。具体而言，我们在这些检测器的骨干网络的最后一层之后，例如 VGGNet-16 [77] 的 *conv5_3* 和 ResNet-50 [63] 的 *res5c* 之后，添加了与我们的 SymFormer 中使用的相同的图像分类头。图像分类头学习将 CXR 图像分类为三类：健康、患病但非结核病、结核病，而结核检测头学习以两类检测结核区域：活动性结核病和陈旧性结核病。现有检测器的训练遵循 §4.4中描述的两阶段训练方式。

### 5.3 评估指标

**CXR 图像分类.** 我们继续介绍 CTD 任务的评估指标。在 CXR 图像分类中，目标是将每个 CXR 图像分类为三类之一：健康、患病但非结核病、结核病。为了评估分类结果，我们使用以下评估指标：

- 准确率（Accuracy）：此指标测量所有三个类别被正确分类的 CXR 图像的百分比。
- 曲线下面积（Area Under Curve, AUC）：AUC 计算了接受者操作特性（Receiver Operating Characteristic, ROC）曲线下的面积。ROC 曲线绘制了结核病类别的真正例率（True Positive Rate）与假正例率（False Positive Rate）之间的关系。
- 灵敏度（Sensitivity）：灵敏度量化了被准确识别为结核病的结核病病例的百分比。它代表了结核病类别的召回率（Recall）。
- 特异度（Specificity）：特异性确定了被准确识别为非结核病的非结核病病例的百分比，包括健康和患病但非结核病的类别。它代表了非结核病类别的召回率。
- 平均精度（Average Precision, AP）：AP 计算每个类别的精度，并取所有类别的平均值。它提供了精度的综合度量。
- 平均召回率（Average Recall, AR）：AR 计算每个类别的召回率，并对所有类别的值取平均。它提供了召回率的综合度量。
- 混淆矩阵（Confusion matrix）：混淆矩阵报告了真正例（True Positives, TP）、真负例（True Negatives, TN）、假正例（False Positives, FP）和假负例（False Negatives, FN）的数量。为了便于观察，我们报告了相对于测试 CXR 图像的总数的 TP、TN、FP 和 FN 的比率。
- $F_1$ 分数（$F_1$ score）：此指标是精度和召回率的调和平均值，因此对两者进行平衡的表示。可以通过公式 $2\mathrm{TP}/(2\mathrm{TP}+\mathrm{FP}+\mathrm{FN})$ 计算。

这些指标使得能够从各个角度评估 CXR 图像分类的质量。

**结核病感染区域检测.** 对于结核病检测的评估，我们使用 COCO 基准 [78] 中提出的边界框度量指标—-平均精度（$\mathrm{AP}^{\mathrm{bb}}$）。$\mathrm{AP}^{\mathrm{bb}}$ 广泛用作计算机视觉社区中的主要检测指标 [39], [70], [74], [79]。默认的 $\mathrm{AP}^{\mathrm{bb}}$ 通过在从 0.5 到 0.95 的 IoU（Intersection-over-Union）阈值范围内以 0.05 的步长进行平均计算。此外，我们报告 $\mathrm{AP}^{\mathrm{bb}}_{50}$，它表示 IoU 阈值为 0.5 时的 $\mathrm{AP}^{\mathrm{bb}}$。为了提供对不同类型结核病的检测性能的理解，我们分别为活动性结核病和陈旧性结核病提供评估结果，不包括不确定的结核病 CXR 图像。我们还报告类别无关的结核病检测结果，其中忽略了结核病类别，以描述所有结核区域的检测结果。在这种情况下，包括了不确定的结核病 CXR 图像。此外，我们引入了两种评估模式：i）使用 TBX11K 测试集中的所有 CXR 图像，ii）仅考虑 TBX11K 测试集中的结核病 CXR 图像。通过使用这些指标，我们可以从各个有用的角度全面分析 CTD 系统的性能。

## 6 实验结果

在本节中，我们首先在 §6.1中介绍了 CXR 图像分类结果，然后是 §6.2中的结核病感染区域检测结果。随后，在 §6.3中，我们进行了消融研究，以更好地了解所提出的 SymFormer。最后，我们在 §6.4中可视化了检测结果和学到的深度特征。

### 6.1 CXR 图像分类

我们总结了在 TBX11K 测试集上进行的 CXR 图像分类的评估结果，见表3和表4。所有方法都采用来自 ImageNet [33] 的预训练模型进行初始化。我们报告了集成了 RetinaNet [70] 和 Deformable DETR [69] 的 SymFormer 的结果。从表3和表4都可以看出，将 SymFormer 集成到 RetinaNet [70] 和 Deformable DETR [69] 中显著提升了它们的性能。SymFormer 与 Deformable DETR 一起实现了 97.3% 的特异度，表明 100 个非结核病 CXR 图像中有 2.7 个会被错误分类为结核病。我们采用的默认模型是 SymFormer 与 RetinaNet 的结合，其性能略低于 SymFormer 与 Deformable DETR，但在目标检测方面的性能明显优于后者，如 §6.2所示。此外，就准确率而言，所有方法都大大优于在 §3.2中达到的 84.8% 的放射科医生。这强调了基于深度学习的 CTD 作为一个研究领域具有很大的潜力。

在表3中，我们观察到 SymFormer 与基线模型在灵敏度方面的差异要比在特异度方面的差异小得多。特别是，Faster R-CNN [75] 实现了令人印象深刻的高灵敏度，达到了 91.2%，但在其他性能指标方面远远落后于 SymFormer。为了解释这一现象，我们参考表4，发现基线模型倾向于做更多的阳性预测（TP + FP）和更少的阴性预测（TN + FN）。简而言之，基线模型倾向于将测试 CXR 图像分类为阳性，可能是因为它们学习高质量结核病相关特征的能力有限。当我们在不重

表 3
**在 TBX11K 测试数据上，根据准确率、AUC、灵敏度、特异度、AP 和 AR 的 CXR 图像分类结果（%）。“骨干网络”列指示了使用的骨干网络。**

| 方法 | 骨干网络 | Accuracy | AUC (TB) | Sensitivity | Specificity | Ave. Prec. (AP) | Ave. Rec. (AR) |
|---|---|---|---|---|---|---|---|
| SSD [73] | VGGNet-16 | 84.7 | 93.0 | 78.1 | 89.4 | 82.1 | 83.8 |
| RetinaNet [70] | ResNet-50 w/ FPN | 87.4 | 91.8 | 81.6 | 89.8 | 84.8 | 86.8 |
| Faster R-CNN [75] | ResNet-50 w/ FPN | 89.7 | 93.6 | 91.2 | 89.9 | 87.7 | 90.5 |
| FCOS [74] | ResNet-50 w/ FPN | 88.9 | 92.4 | 87.3 | 89.9 | 86.6 | 89.2 |
| Deformable DETR [69] | ResNet-50 w/ FPN | 91.3 | 97.6 | 89.2 | 95.3 | 89.8 | 91.0 |
| SymFormer w/ Deformable DETR | ResNet-50 w/ FPN | 94.3 | 98.5 | 87.3 | **97.3** | 93.2 | 93.2 |
| SymFormer w/ RetinaNet | ResNet-50 w/ FPN | 94.5 | 98.9 | 91.0 | 96.8 | 93.3 | 94.0 |
| SymFormer w/ RetinaNet | P2T-Small w/ FPN | **94.6** | **99.1** | **92.1** | 96.7 | **93.4** | **94.2** |

表 4
**在 TBX11K 测试数据上，根据 $F_1$ 分数和混淆矩阵的 CXR 图像分类结果（%），以及每个模型的 FLOPs 数量、参数数量和 FPS。“#Total”表示测试 CXR 图像的总数。我们在单个 TITAN XP GPU 上测试 FPS。对于真值，阳性（TP + FN）比例为 19.6%，阴性（TN + FP）比例为 80.4%。**

| 方法 | 骨干网络 | FLOPs 数量 | 参数量 | FPS | $F_1$ 分数 ↑ | $\frac{\text{TP}}{\#\text{Total}}$ ↑ | $\frac{\text{TN}}{\#\text{Total}}$ ↑ | $\frac{\text{FP}}{\#\text{Total}}$ ↓ | $\frac{\text{FN}}{\#\text{Total}}$ ↓ |
|---|---|---|---|---|---|---|---|---|---|
| SSD [73] | VGGNet-16 | 90.58 | 38.69 | 32.9 | 70.5 | 15.3 | 71.9 | 8.5 | 4.3 |
| RetinaNet [70] | ResNet-50 w/ FPN | 55.41 | 48.97 | 35.3 | 73.1 | 16.0 | 72.2 | 8.2 | 3.6 |
| Faster R-CNN [75] | ResNet-50 w/ FPN | 66.27 | 53.98 | 30.3 | 78.5 | 17.9 | 72.3 | 8.1 | 1.7 |
| FCOS [74] | ResNet-50 w/ FPN | 53.33 | 44.69 | 39.9 | 76.3 | 17.1 | 72.3 | 8.1 | 2.5 |
| Deformable DETR [69] | ResNet-50 w/ FPN | 54.07 | 52.67 | 23.0 | 85.6 | 17.5 | 76.6 | 3.8 | 2.1 |
| SymFormer w/ Deformable DETR | ResNet-50 w/ FPN | 54.08 | 52.69 | 22.5 | 87.9 | 17.1 | **78.2** | **2.2** | 2.5 |
| SymFormer w/ RetinaNet | ResNet-50 w/ FPN | 59.14 | 50.03 | 24.3 | 89.0 | 17.8 | 77.8 | 2.6 | 1.8 |
| SymFormer w/ RetinaNet | P2T-Small w/ FPN | 55.46 | 45.10 | 17.9 | **89.6** | **18.1** | 77.7 | 2.7 | **1.5** |

新训练的情况下在其他公共数据集上评估所有模型，如表5和表6所示，我们可以看到基线模型甚至实现了 100.0% 的灵敏度和 0 的特异度。这进一步证实了我们的假设，即基线模型倾向于将 CXR 图像分类为阳性。考虑到这一观点，表4中的 $F_1$ 分数更好地代表了模型的整体性能，因为它对准确率和召回率进行了平衡组合。

在表4中，我们还报告了每个模型的浮点运算次数（Floating-Point Operations, FLOPs）、参数数量和每秒帧数（Frames Per Second, FPS）。通过比较原始的和经 SymFormer 增强的 Deformable DETR [69]，我们可以看到 SymFormer 在 FLOPs、参数和运行速度方面与 Deformable DETR 相似。这是显而易见的，因为 SymFormer 对 Deformable DETR 引入的计算影响可以忽略不计。当与使用 ResNet-50 [63] 骨干的 RetinaNet [70] 集成时，SymFormer 实现了 24.3 fps，使其成为可以在现实场景中部署的选择。如果有额外的计算资源，SymFormer 还可以利用 P2T-Small [39] 作为骨干，提供增强的诊断性能和 17.9 fps 的速度。

此外，我们使用上述训练过的模型在未进行微调的情况下在公共数据集上评估 CXR 图像分类。在表5中呈现了 DA+DB 测试数据 [7] 上的结果，在表6中呈现了 MC+Shenzhen 测试数据 [26] 上的结果。值得注意的是，SSD [73]、RetinaNet [70]、Faster R-CNN [75] 和 FCOS [74] 实现了接近 100.0% 的灵敏度和大约 0 的特异度。正如前面讨论的，这表明基线模型难以学习用于 CTD 的强大特征表示，经常将测试 CXR 图像错误地分类为结核病病例。Deformable DETR [69] 在公共数据集上展现了一定的泛化能力，但与提出的 SymFormer 相比仍有差距。SymFormer 的强大性能突显了其出色的泛化能力。

### 6.2 结核病感染区域检测

我们继续展示结核病感染区域检测的结果。如 §5.3中所讨论的，我们报告了整个 TBX11K 测试集和仅包含结核病 CXR 图像子集的性能。仅使用结核病 CXR 图像进行性能评估可以对检测结果进行精确的分析，因为非结核病 CXR 图像不包含结核病感染区域。相反，使用所有 CXR 图像进行评估会考虑非结核病 CXR 图像中假阳性的影响。当使用所有 CXR 图像进行评估时，为了确保准确，我们舍弃了由 CXR 图像分类头分类为非结核病的 CXR 图像中的所有预测框。然而，值得注意的是，当仅使用结核病 CXR 图像进行评估时，这个过滤过程是不适用的。

结核病感染区域检测的结果在表7中呈现。显然，SymFormer w/ Deformable DETR 和 SymFormer w/ RetinaNet 在各自的基础方法 Deformable DETR [69] 和 RetinaNet [70] 的基础上均取得了显著的提升。有趣的是，SymFormer w/ RetinaNet 在性能上优于 SymFormer w/ Deformable DETR，表明 SymFormer 更适合与 RetinaNet 框架集成。因此，我们

表 5
**在 DA+DB 测试数据 [7] 上，根据准确率、AUC、灵敏度、特异度、AP 和 AR 的 CXR 图像分类结果（%）。**

| 方法 | 骨干网络 | Accuracy | AUC (TB) | Sensitivity | Specificity | Ave. Prec. (AP) | Ave. Rec. (AR) |
|---|---|---|---|---|---|---|---|
| SSD [73] | VGGNet-16 | 51.0 | 53.8 | **100.0** | 1.9 | 75.3 | 51.0 |
| RetinaNet [70] | ResNet-50 w/ FPN | 50.0 | 50.0 | **100.0** | 0.0 | 25.0 | 50.0 |
| Faster R-CNN [75] | ResNet-50 w/ FPN | 50.0 | 51.9 | **100.0** | 0.0 | 25.0 | 50.0 |
| FCOS [74] | ResNet-50 w/ FPN | 50.0 | 52.1 | **100.0** | 0.0 | 25.0 | 50.0 |
| Deformable DETR [69] | ResNet-50 w/ FPN | 68.6 | 69.7 | 84.3 | 52.9 | 70.7 | 68.6 |
| SymFormer w/ Deformable DETR | ResNet-50 w/ FPN | 82.4 | 78.4 | 86.3 | 78.4 | 82.6 | 82.4 |
| SymFormer w/ RetinaNet | ResNet-50 w/ FPN | 78.4 | 74.7 | 90.2 | 66.7 | 80.1 | 78.4 |
| SymFormer w/ RetinaNet | P2T-Small w/ FPN | **84.3** | **89.4** | 84.3 | **84.3** | **84.3** | **84.3** |

表 6
**在 MC+Shenzhen 测试数据 [26] 上，根据准确率、AUC、灵敏度、特异度、AP 和 AR 的 CXR 图像分类结果（%）。**

| 方法 | 骨干网络 | Accuracy | AUC (TB) | Sensitivity | Specificity | Ave. Prec. (AP) | Ave. Rec. (AR) |
|---|---|---|---|---|---|---|---|
| SSD [73] | VGGNet-16 | 50.8 | 50.4 | **100.0** | 3.4 | 74.9 | 51.7 |
| RetinaNet [70] | ResNet-50 w/ FPN | 49.3 | 49.7 | **100.0** | 0.5 | 74.6 | 50.3 |
| Faster R-CNN [75] | ResNet-50 w/ FPN | 49.0 | 49.5 | **100.0** | 0.0 | 24.5 | 50.0 |
| FCOS [74] | ResNet-50 w/ FPN | 48.8 | 49.0 | 99.5 | 0.0 | 24.4 | 49.7 |
| Deformable DETR [69] | ResNet-50 w/ FPN | 81.3 | 83.5 | 92.9 | 70.1 | 83.0 | 81.5 |
| SymFormer w/ Deformable DETR | ResNet-50 w/ FPN | 82.0 | 84.7 | 89.3 | 75.0 | 82.7 | 82.1 |
| SymFormer w/ RetinaNet | ResNet-50 w/ FPN | 82.8 | 86.3 | 91.8 | 74.0 | 83.8 | 82.9 |
| SymFormer w/ RetinaNet | P2T-Small w/ FPN | **85.8** | **87.4** | 93.4 | **78.4** | **86.6** | **85.9** |

表 7
**在我们的 TBX11K 测试集上的结核病感染区域检测结果（%）。“测试数据”列指定了是使用测试集中的所有 CXR 图像还是仅使用测试集中的结核病 CXR 图像进行评估。“骨干网络”列指示了使用的骨干网络。**

| 方法 | 测试数据 | 骨干网络 | 类别无关的结核 | | 活动性结核 | | 陈旧性结核 | |
|---|---|---|---|---|---|---|---|---|
| | | | $AP^{bb}_{50}$ | $AP^{bb}$ | $AP^{bb}_{50}$ | $AP^{bb}$ | $AP^{bb}_{50}$ | $AP^{bb}$ |
| SSD [73] | ALL | VGGNet-16 | 52.3 | 22.6 | 50.5 | 22.8 | 8.1 | 3.2 |
| RetinaNet [70] | | ResNet-50 w/ FPN | 52.1 | 22.2 | 45.4 | 19.6 | 6.2 | 2.4 |
| Faster R-CNN [75] | | ResNet-50 w/ FPN | 57.3 | 22.7 | 53.3 | 21.9 | 9.6 | 2.9 |
| FCOS [74] | | ResNet-50 w/ FPN | 46.6 | 18.9 | 40.3 | 16.8 | 6.2 | 2.1 |
| Deformable DETR [69] | | ResNet-50 w/ FPN | 51.7 | 22.0 | 48.9 | 21.2 | 7.1 | 1.9 |
| SymFormer w/ Deformable DETR | | ResNet-50 w/ FPN | 57.0 | 23.3 | 52.1 | 22.7 | 7.1 | 2.0 |
| SymFormer w/ RetinaNet | | ResNet-50 w/ FPN | 68.0 | 29.5 | 62.0 | **27.3** | **13.3** | **4.4** |
| SymFormer w/ RetinaNet | | P2T-Small w/ FPN | **70.4** | **30.0** | **63.6** | 26.9 | 11.4 | 4.3 |
| SSD [73] | Only TB | VGGNet-16 | 68.3 | 28.7 | 63.7 | 28.0 | 10.7 | 4.0 |
| RetinaNet [70] | | ResNet-50 w/ FPN | 69.4 | 28.3 | 61.5 | 25.3 | 10.2 | 4.1 |
| Faster R-CNN [75] | | ResNet-50 w/ FPN | 63.4 | 24.6 | 58.7 | 23.7 | 9.6 | 2.8 |
| FCOS [74] | | ResNet-50 w/ FPN | 56.3 | 22.5 | 47.9 | 19.8 | 7.4 | 2.4 |
| Deformable DETR [69] | | ResNet-50 w/ FPN | 57.4 | 24.2 | 54.5 | 23.5 | 7.6 | 2.3 |
| SymFormer w/ Deformable DETR | | ResNet-50 w/ FPN | 60.8 | 24.5 | 55.2 | 23.8 | 9.2 | 2.6 |
| SymFormer w/ RetinaNet | | ResNet-50 w/ FPN | 73.4 | 31.5 | 67.1 | **29.2** | **14.7** | **4.8** |
| SymFormer w/ RetinaNet | | P2T-Small w/ FPN | **75.7** | **32.1** | **68.9** | 28.9 | 13.0 | 4.7 |

选择 SymFormer w/ RetinaNet 作为我们 CTD 的默认模型。值得注意的是，所有方法在准确检测陈旧性结核区域方面都存在困难。然而，与活动性结核病相比，基于类别无关的结核病评估的结果更好，表明许多陈旧性结核区域被正确定位，但被错误地分类为活动性结核。我们将这归因于 TBX11K 数据集中陈旧性结核病 CXR 图像的数量有限，与包含活动性结核的 924 张 CXR 图像相比，只有 212 张 CXR 图像包含陈旧性结核。因此，未来的研究应该解决这个数据不平衡的问题，

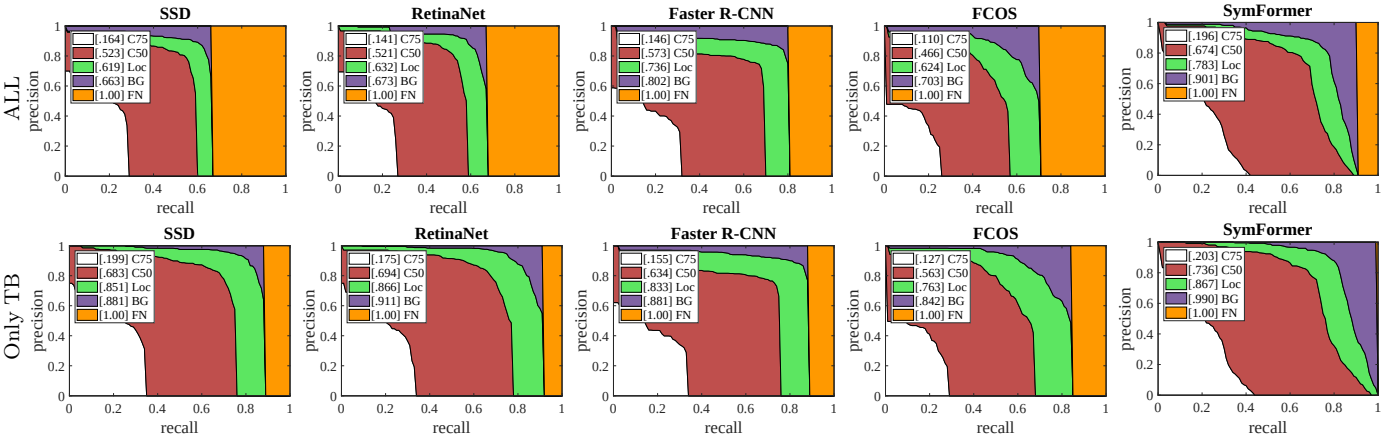


**图 5. 使用基线模型和 SymFormer w/ RetinaNet 进行类别无关的结核病感染区域检测的错误分析。**第一行是使用所有 CXR 图像进行评估，而第二行仅使用结核病 CXR 图像。C50/C75：IoU 阈值为 0.5/0.75 下的 PR 曲线；Loc：IoU 阈值为 0.1 下的 PR 曲线；BG：去除背景假阳性；FN：去除由未检测到的目标引起的其他错误（假阴性）。在所有指标上，SymFormer 在很大程度上优于其他方法，例如，仅使用结核病 CXR 图像时获得了惊人的 99% 的 BG 分数。

并专注于改进陈旧性结核区域的检测。此外，我们观察到在 $AP^{bb}_{50}$ 方面的性能通常优于 $AP^{bb}$。这表明，虽然检测模型能够识别目标区域，但其定位通常不太精确。我们认为，定位结核边界框区域与定位自然中物体的区域存在显著差异。即使是经验丰富的放射科医生也发现精确确定结核区域是具有挑战性的。因此，$AP^{bb}_{50}$ 比 $AP^{bb}$ 更为重要，因为与目标结核区域具有 IoU 为 0.5 的预测框足以帮助放射科医生识别结核病感染区域。

在图5中，我们呈现了用于检测错误分析的精度-召回率（Precision-Recall, PR）曲线，关注类别无关的结核病检测。显然，当 IoU 阈值从 0.75 过渡到 0.5 时，所有方法都表现出显著的提高。这表明，由于它们有限的目标定位能力，所有方法在较高的 IoU 阈值下的性能都特别受到挑战。将使用所有 CXR 图像的结果与仅使用结核病 CXR 图像的结果进行比较，我们观察到在使用所有 CXR 图像进行评估时，“FN”（假阴性）区域较大。这表明，基于图像分类的过滤过程忽略了许多正确检测到的结核病区域，尽管它在提高整体检测性能方面是有效的。重要的是，SymFormer 的 “FN” 区域明显小于其他方法，突显了其更强大的检测较少假阴性的能力。无论是使用所有 CXR 图像还是仅使用结核病 CXR 图像，SymFormer 在 IoU 阈值为 0.75、0.5 和 0.1 时始终展现出更高的 PR 曲线。通过综合考虑图像分类和结核病感染区域检测的结果，我们可以自信地得出结论，本文所提出的 SymFormer 在 CTD 领域取得了最先进的性能，为将来的研究提供了强大的基线。

### 6.3 消融研究

在这一部分，我们首先进行了消融研究，以评估所提出模块的有效性。具体而言，我们使用 TBX11K 数据集的训练集训练模型，并在验证集上进行评估。结果呈现在表8中。基线模型是 RetinaNet [70]，对应于表8中的第一个模型，不包含任何注意力或位置编码。术语 “vanilla attention” 指的是 Deformable DETR [69] 中使用的可变形注意力。我们使用了公开的绝对位置编码 [36], [37]（如公式1中所述）和相对位置编码 [80] 的实现。根据公式2，SPE 的默认版本将位置编码的右侧转移到左侧。在这里，我们还评估了将左侧转移到右侧时的性能。

**表 8**
**在我们 TBX11K 验证集上进行的结核病感染区域检测的消融研究。**我们只使用结核病 CXR 图像来评估类别无关的结核区域检测。“SPE 对称方式” 指示 SPE 是否将位置编码的右侧转移到左侧，或者相反。APE：绝对位置编码；RPE：相对位置编码。

| 注意力 | 位置编码 | SPE 对称方式 | $AP^{bb}_{50}$ | $AP^{bb}$ |
|---|---|---|---|---|
| No | No | - | 72.7 | 31.0 |
| Vanilla | APE | - | 73.4 | 30.6 |
| Vanilla | RPE | - | 72.7 | 29.7 |
| Vanilla | SPE w/o STN | left → right | 74.0 | 30.5 |
| Vanilla | SPE w/o STN | right → left | 74.3 | 30.8 |
| Vanilla | SPE | left → right | 75.1 | 30.4 |
| Vanilla | SPE | right → left | 75.7 | 29.6 |
| SymAttention | APE | - | 74.9 | 30.0 |
| SymAttention | RPE | - | 73.6 | 29.1 |
| SymAttention | SPE w/o STN | left → right | 75.3 | 31.4 |
| SymAttention | SPE w/o STN | right → left | 75.5 | 30.7 |
| SymAttention | SPE | left → right | 76.3 | 30.9 |
| SymAttention | SPE | right → left | **76.6** | **31.7** |

根据 §6.2中的讨论，$AP^{bb}_{50}$ 指标被认为足以衡量模型在帮助放射科医生识别结核病感染区域方面的有效性。正如从表8中可以看出的，相对位置编码相对于绝对位置编码表现较差，所以我们构建 SPE 时使用了绝对位置编码。此外，将绝对位置编码和任何形式的注意力添加到 RetinaNet [70] 中均显著提高了检测性能。此外，对于所有类型的位置编码，我们提出的 SymAttention 都始终优于可变形注意力，展示了其在学习 CTD 的独特表示方面的优越性。值得注意的是，即使没有 STN，所提出的 SPE 在性能上也始终优于绝对位置编码和

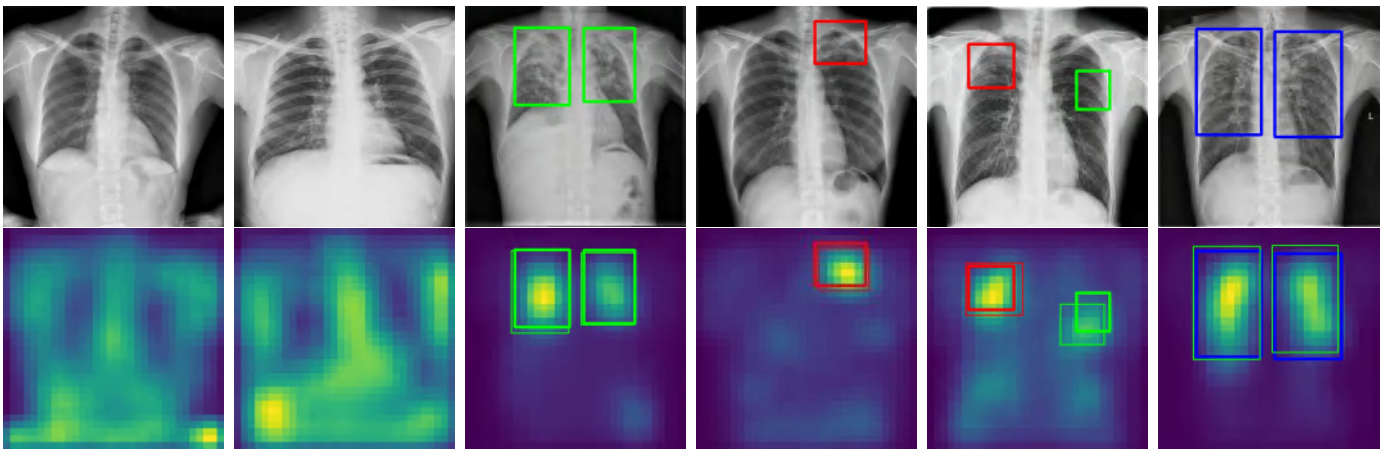


**图 6. 可视化 SymFormer w/ RetinaNet 从 CXR 图像中学到的深度特征。**我们从 TBX11K 测试集中为表2中提到的每个类别随机选择一张 CXR 图像。在每个示例中，活动性结核、陈旧性结核和不确定结核的感染区域分别由绿色、红色和蓝色的框表示。真值边界框显示为粗线，而检测到的边界框显示为细线。

**表 9**
**4 折交叉验证的评估结果（%）。**所使用的模型是我们的 SymFormer w/ RetinaNet，使用 ResNet-50 w/ FPN 作为骨干网络。我们将 TBX11K `trainval` 集分成 4 个折，每个折具有相似的类别分布。"#Total" 表示每个折中测试 CXR 图像的总数。我们只使用结核病 CXR 图像来评估类别无关的结核区域检测。

| 折 | Accuracy | AUC | Sensitivity | Specificity | AP | AR | $F_1$ 分数 | $\frac{\text{TP}}{\#\text{Total}}$ | $\frac{\text{TN}}{\#\text{Total}}$ | $\frac{\text{FP}}{\#\text{Total}}$ | $\frac{\text{FN}}{\#\text{Total}}$ | $\text{AP}^{\text{bb}}_{50}$ | $\text{AP}^{\text{bb}}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 94.7 | 98.9 | 92.2 | 97.5 | 91.1 | 94.6 | 87.6 | 11.0 | 85.8 | 2.2 | 0.9 | 74.7 | 31.7 |
| 2 | 95.2 | 98.9 | 92.6 | 97.6 | 91.7 | 95.0 | 88.1 | 11.1 | 86.0 | 2.1 | 0.9 | 75.2 | 30.4 |
| 3 | 94.6 | 99.1 | 92.9 | 96.9 | 90.5 | 94.6 | 86.4 | 11.1 | 85.3 | 2.7 | 0.8 | 74.1 | 29.5 |
| 4 | 95.1 | 99.3 | 92.9 | 97.3 | 91.2 | 94.7 | 87.4 | 11.1 | 85.7 | 2.4 | 0.8 | 75.4 | 33.3 |

相对位置编码。STN 的引入进一步提高了 SPE 的性能，证实了其有效性。因此，我们对 CTD 中的对称异常搜索的研究取得了成功的结果。此外，我们可以观察到，在 SPE 的对称方式方面，从右侧到左侧的位置编码转换，相较于从左侧到右侧，略微优越。因此，默认情况下我们将位置编码从右侧转换到左侧。

为了展示所提出的 SymFormer 的稳健性，我们进行了 4 折交叉验证。我们将 TBX11K 的 `trainval` 数据分成四个折，确保每个折具有相似的类别分布。在每次试验中，我们在这四个折中的三个上训练 SymFormer，并在剩余的一个折上评估其性能。我们将四次试验的评估结果显示在表9中。可以看出，这些不同试验之间的结果非常一致，证实了 SymFormer 的稳健性。

### 6.4 可视化

为了深入了解深度神经网络在 CXR 图像上的学习过程，我们以 1/32 的分辨率可视化了 SymFormer w/ RetinaNet 的特征图。为实现这一目标，我们使用主成分分析（Principal Component Analysis, PCA）将特征图的通道减少到一个单通道。然后，将生成的单通道图转换为热图以进行可视化。学习到的特征的可视化结果以及相应的检测结果呈现在图6中。经过分析，我们观察到，健康病例的可视化呈现不规则的特征模式，表明不存在显著的异常。相比之下，患病但非结核病病例的可视化显示出一些可辨别的亮点，可能表示存在病变。对于结核病病例，可视化图中的亮点与标注的肺结核感染区域很好地对齐，从而表明所提出的 SymFormer 在学习肺结核感染区域的深度特征方面是有效的。此外，在图7中，我们提供了所提出的 SymFormer 与基线模型在结核病感染区域检测方面的定性比较。明显地，SymFormer 始终呈现出更为出色的定性检测结果。

## 7 总结

早期诊断在有效治疗和预防世界范围内流行的结核病中起着至关重要的作用。然而，结核病的诊断仍然是一个重大挑战，特别是在资源有限的社区和发展中国家。传统的结核病金标准检测方法需要 BSL-3 实验室，并且是一个耗时的过程，需要数月才能提供明确的结果，这在许多情境中都是不切实际的。深度学习在各个领域取得了令人期待的进展，促使研究人员探索其在计算机辅助结核病诊断中的潜力。然而，缺乏带有边界框标注的数据阻碍了深度学习在这一领域的进展。为了解决这一限制，我们引入了 TBX11K，这是一个带有边界框标注的大规模结核病数据集。这个数据集不仅有助于为 CTD

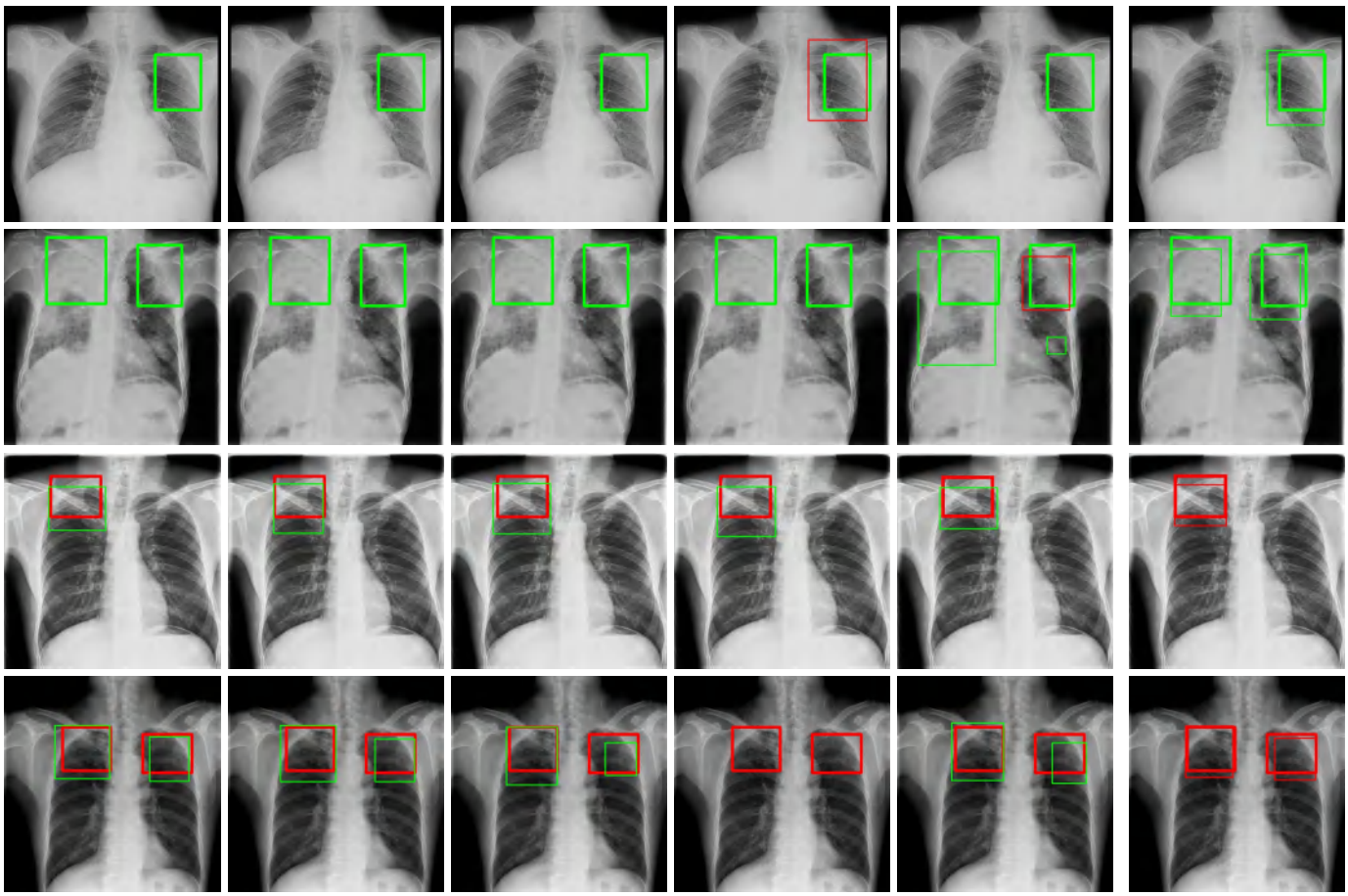


**图 7. 所提出的 SymFormer 与基线方法的定性比较。在每个示例中，活动性结核和陈旧性结核的感染区域分别用绿色框和红色框表示。真值边界框用粗线显示，而检测到的边界框用细线显示。对于所有示例，SymFormer 能够检测到所有的结核病感染区域并正确预测类别。**

的深度神经网络训练提供支持，还是第一个专门设计用于结核感染区域检测的数据集。

除了数据集之外，我们提出了一个名为 SymFormer 的简单而有效的框架，用于同时进行 CXR 图像分类和结核感染区域检测。利用 CXR 图像固有的双侧对称性属性，SymFormer 引入了对称搜索注意力（SymAttention）以提取独特的特征表示。由于 CXR 图像可能不呈现严格的对称性，我们引入了对称位置编码（SPE），通过特征校准来增强 SymAttention 的性能。此外，为了为 CTD 研究提供一个基准，我们引入了评估指标，评估从现有目标检测器改进得到的基线模型，并启动了一个在线挑战。我们的实验证明，SymFormer 在 TBX11K 数据集上取得了最先进的性能，使其成为未来研究的强有力的基准。本研究中引入的 TBX11K 数据集、SymFormer 方法以及 CTD 基准预计将显著推动 CTD 领域的研究，最终有助于改善全球结核病的检测和管理。

## 致谢

本工作部分得到了中国国家重点研发计划（No. 2021YFB3100800）和中国国家自然科学基金（No. 62376283）的支持。

## 参考文献

[1] Y. Liu, Y.-H. Wu, S.-C. Zhang, L. Liu, M. Wu, and M.-M. Cheng, "Revisiting computer-aided tuberculosis diagnosis," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2023.

[2] Y. Liu, Y.-H. Wu, Y. Ban, H. Wang, and M.-M. Cheng, "Rethinking computer-aided tuberculosis diagnosis," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 2646–2655.

[3] W. H. Organization, "Global tuberculosis report 2015," https://apps.who.int/iris/handle/10665/191102, 2015.

[4] ——, "Global tuberculosis report 2017," https://apps.who.int/iris/handle/10665/259366, 2017.

[5] ——, "Global tuberculosis report 2020," https://www.who.int/publications/i/item/9789240013131, 2020.

[6] S. Jaeger, A. Karargyris, S. Candemir, L. Folio, J. Siegelman, F. Callaghan, Z. Xue, K. Palaniappan, R. K. Singh, S. Antani *et al.*, "Automatic tuberculosis screening using chest radiographs," *IEEE Trans. Med. Imaging*, vol. 33, no. 2, pp. 233–245, 2013.

[7] A. Chauhan, D. Chauhan, and C. Rout, "Role of gist and phog features in computer-aided diagnosis of tuberculosis without segmentation," *PloS One*, vol. 9, no. 11, p. e112980, 2014.

[8] S. Hwang, H.-E. Kim, J. Jeong, and H.-J. Kim, "A novel approach for tuberculosis screening based on deep convolutional neural networks," in *Medical Imaging: Computer-Aided Diagno-*

*sis*, vol. 9785. International Society for Optics and Photonics, 2016, p. 97852W.

[9] N. R. Gandhi, P. Nunn, K. Dheda, H. S. Schaaf, M. Zignol, D. Van Soolingen, P. Jensen, and J. Bayona, "Multidrug-resistant and extensively drug-resistant tuberculosis: a threat to global control of tuberculosis," *The Lancet*, vol. 375, no. 9728, pp. 1830–1843, 2010.

[10] U. Lopes and J. F. Valiati, "Pre-trained convolutional neural networks as feature extractors for tuberculosis detection," *Computers in Biology and Medicine*, vol. 89, pp. 135–143, 2017.

[11] S. Candemir, S. Jaeger, K. Palaniappan, J. P. Musco, R. K. Singh, Z. Xue, A. Karargyris, S. Antani, G. Thoma, and C. J. McDonald, "Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration," *IEEE Trans. Med. Imaging*, vol. 33, no. 2, pp. 577–590, 2013.

[12] P. Andersen, M. Munk, J. Pollock, and T. Doherty, "Specific immune-based diagnosis of tuberculosis," *The Lancet*, vol. 356, no. 9235, pp. 1099–1104, 2000.

[13] A. Bekmurzayeva, M. Sypabekova, and D. Kanayeva, "Tuberculosis diagnosis using immunodominant, secreted antigens of mycobacterium tuberculosis," *Tuberculosis*, vol. 93, no. 4, pp. 381–388, 2013.

[14] W. H. Organization *et al.*, "Chest radiography in tuberculosis detection: Summary of current WHO recommendations and guidance on programmatic approaches," World Health Organization, Tech. Rep., 2016.

[15] M. Van Cleeff, L. Kivihya-Ndugga, H. Meme, J. Odhiambo, and P. Klatser, "The role and performance of chest X-ray for the diagnosis of tuberculosis: A cost-effectiveness analysis in nairobi, kenya," *BMC Infectious Diseases*, vol. 5, no. 1, p. 111, 2005.

[16] A. Konstantinos, "Testing for tuberculosis," *Australian Prescriber*, vol. 33, no. 1, pp. 12–18, 2010.

[17] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Adv. Neural Inform. Process. Syst.*, 2014, pp. 1988–1996.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034.

[19] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Int. Conf. Comput. Vis.*, 2017, pp. 2961–2969.

[20] T. Xu, X.-F. Zhu, and X.-J. Wu, "Learning spatio-temporal discriminative model for affine subspace based visual object tracking," *Visual Intelligence*, vol. 1, no. 1, p. 4, 2023.

[21] Y. Liu, M.-M. Cheng, X. Hu, J.-W. Bian, L. Zhang, X. Bai, and J. Tang, "Richer convolutional features for edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1939–1946, 2019.

[22] Y. Liu, M.-M. Cheng, D.-P. Fan, L. Zhang, J.-W. Bian, and D. Tao, "Semantic edge detection with diverse deep supervision," *Int. J. Comput. Vis.*, vol. 130, no. 1, pp. 179–198, 2022.

[23] J. Li, Z. Li, L. Wei, and X. Zhang, "Machine learning in lung cancer radiomics," *Machine Intelligence Research*, 2023.

[24] Y. Qiu, F. Lin, W. Chen, and M. Xu, "Pre-training in medical data: A survey," *Machine Intelligence Research*, vol. 20, no. 2, pp. 147–179, 2023.

[25] W.-C. Wang, E. Ahn, D. Feng, and J. Kim, "A review of predictive and contrastive self-supervised learning for medical images," *Machine Intelligence Research*, vol. 20, no. 4, pp. 483–513, 2023.

[26] S. Jaeger, S. Candemir, S. Antani, Y.-X. J. Wáng, P.-X. Lu, and G. Thoma, "Two public chest X-ray datasets for computer-aided screening of pulmonary diseases," *Quantitative Imaging in Medicine and Surgery*, vol. 4, no. 6, pp. 475–477, 2014.

[27] Y. Boykov and G. Funka-Lea, "Graph cuts and efficient N-D image segmentation," *Int. J. Comput. Vis.*, vol. 70, no. 2, pp. 109–131, 2006.

[28] A. Oliva and A. Torralba, "Building the gist of a scene: The role of global image features in recognition," *Progress in Brain Research*, vol. 155, pp. 23–36, 2006.

[29] A. Bosch, A. Zisserman, and X. Munoz, "Representing shape with a spatial pyramid kernel," in *ACM International Conference on Image and Video Retrieval*. ACM, 2007, pp. 401–408.

[30] M.-M. Cheng, Y. Liu, Q. Hou, J. Bian, P. Torr, S.-M. Hu, and Z. Tu, "HFS: Hierarchical feature selection for efficient image segmentation," in *Eur. Conf. Comput. Vis.*, 2016, pp. 867–882.

[31] Y. Liu, P.-T. Jiang, V. Petrosyan, S.-J. Li, J. Bian, L. Zhang, and M.-M. Cheng, "DEL: Deep embedding learning for efficient image segmentation," in *Int. Joint Conf. Artif. Intell.*, 2018, pp. 864–870.

[32] A. Karargyris, J. Siegelman, D. Tzortzis, S. Jaeger, S. Candemir, Z. Xue, K. Santosh, S. Vajda, S. Antani, L. Folio *et al.*, "Combination of texture and shape features to detect pulmonary abnormalities in digital chest X-rays," *International Journal of Computer Assisted Radiology and Surgery*, vol. 11, no. 1, pp. 99–106, 2016.

[33] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2009, pp. 248–255.

[34] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Adv. Neural Inform. Process. Syst.*, 2012, pp. 1097–1105.

[35] P. Lakhani and B. Sundaram, "Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks," *Radiology*, vol. 284, no. 2, pp. 574–582, 2017.

[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Adv. Neural Inform. Process. Syst.*, pp. 5998–6008, 2017.

[37] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Int. Conf. Learn. Represent.*, 2021.

[38] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Int. Conf. Mach. Learn.*, 2021, pp. 10 347–10 357.

[39] Y.-H. Wu, Y. Liu, X. Zhan, and M.-M. Cheng, "P2T: Pyramid pooling transformer for scene understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022.

[40] Y. Zhang, H. Liu, and Q. Hu, "TransFuse: Fusing transformers and CNNs for medical image segmentation," in *Med. Image. Comput. Comput. Assist. Interv.*, 2021, pp. 14–24.

[41] Y. Xie, J. Zhang, C. Shen, and Y. Xia, "CoTr: Efficiently bridging CNN and transformer for 3D medical image segmentation," in *Med. Image. Comput. Comput. Assist. Interv.*, 2021, pp. 171–180.

[42] Y. Ji, R. Zhang, H. Wang, Z. Li, L. Wu, S. Zhang, and P. Luo, "Multi-compound transformer for accurate biomedical image segmentation," in *Med. Image. Comput. Comput. Assist. Interv.*, 2021, pp. 326–336.

[43] Y. Gao, M. Zhou, and D. N. Metaxas, "UTNet: A hybrid transformer architecture for medical image segmentation," in *Med. Image. Comput. Comput. Assist. Interv.*, 2021, pp. 61–71.

[44] R. Tao, W. Liu, and G. Zheng, "Spine-transformers: Vertebra labeling and segmentation in arbitrary field-of-view spine CTs via 3D transformers," *Med. Image Anal.*, vol. 75, p. 102258, 2022.

[45] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. Ji *et al.*, "TransMIL: Transformer based correlated multiple instance learning for whole slide image classification," *Adv. Neural Inform. Process. Syst.*, pp. 2136–2147, 2021.

[46] S. Park, G. Kim, J. Kim, B. Kim, and J. C. Ye, "Federated split task-agnostic vision transformer for COVID-19 CXR diagnosis," *Adv. Neural Inform. Process. Syst.*, pp. 24 617–24 630, 2021.

[47] Y. Mo, C. Han, Y. Liu, M. Liu, Z. Shi, J. Lin, B. Zhao, C. Huang, B. Qiu, Y. Cui *et al.*, "HoVer-Trans: Anatomy-aware HoVer-Transformer for ROI-free breast cancer diagnosis in ultrasound images," *IEEE Trans. Med. Imaging*, vol. 42, no. 6, pp. 1696–1706, 2023.

[48] M. Bhattacharya, S. Jain, and P. Prasanna, "RadioTransformer: A cascaded global-focal transformer for visual attention–guided disease classification," in *Eur. Conf. Comput. Vis.*, 2022, pp. 679–698.

[49] S. Park, G. Kim, Y. Oh, J. B. Seo, S. M. Lee, J. H. Kim, S. Moon, J.-K. Lim, and J. C. Ye, "Multi-task vision transformer using low-level chest X-ray feature corpus for COVID-19 diagnosis and severity quantification," *Med. Image Anal.*, vol. 75, p. 102299, 2022.

[50] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.

[51] Z. Shen, R. Fu, C. Lin, and S. Zheng, "COTR: Convolution in transformer network for end to end polyp detection," in *Int. Conf. on Computer and Communications*, 2021, pp. 1757–1761.

[52] T. S. Mathai, S. Lee, D. C. Elton, T. C. Shen, Y. Peng, Z. Lu, and R. M. Summers, "Lymph node detection in T2 MRI with transformers," in *Medical Imaging 2022: Computer-Aided Diagnosis*, vol. 12033, 2022, pp. 855–859.

[53] H. Li, L. Chen, H. Han, and S. Kevin Zhou, "SATr: Slice attention with transformer for universal lesion detection," in *Med. Image. Comput. Comput. Assist. Interv.*, 2022, pp. 163–174.

[54] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, and H. Fu, "Transformers in medical imaging: A survey," *arXiv preprint arXiv:2201.09873*, 2022.

[55] K. He, C. Gan, Z. Li, I. Rekik, Z. Yin, W. Ji, Y. Gao, Q. Wang, J. Zhang, and D. Shen, "Transformers in medical image analysis," *Intelligent Medicine*, vol. 3, no. 1, pp. 59–78, 2023.

[56] J. Li, J. Chen, Y. Tang, C. Wang, B. A. Landman, and S. K. Zhou, "Transforming medical imaging with transformers? a comparative review of key properties, current progresses, and future perspectives," *Med. Image Anal.*, vol. 85, p. 102762, 2023.

[57] R. Byng-Maddick and M. Noursadeghi, "Does tuberculosis threaten our ageing populations?" *BMC Infectious Diseases*, vol. 16, no. 1, pp. 1–5, 2016.

[58] A. L. García-Basteiro, H. S. Schaaf, R. Diel, and G. B. Migliori, "Adolescents and young adults: a neglected population group for tuberculosis surveillance," *European Respiratory Journal*, vol. 51, no. 2, p. 1800176, 2018.

[59] Z. Dong, Q.-Q. Wang, S.-C. Yu, F. Huang, J.-J. Liu, H.-Y. Yao, and Y.-L. Zhao, "Age-period-cohort analysis of pulmonary tuberculosis reported incidence, china, 2006-2020," *Infectious Diseases of Poverty*, vol. 11, no. 04, pp. 62–71, 2022.

[60] S. Nhamoyebonde and A. Leslie, "Biological differences between the sexes and susceptibility to tuberculosis," *The Journal of Infectious Diseases*, vol. 209, no. suppl_3, pp. S100–S106, 2014.

[61] K. C. Horton, P. MacPherson, R. M. Houben, R. G. White, and E. L. Corbett, "Sex differences in tuberculosis burden and notifications in low-and middle-income countries: A systematic review and meta-analysis," *PLoS Medicine*, vol. 13, no. 9, p. e1002119, 2016.

[62] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 2117–2125.

[63] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 770–778.

[64] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," in *Adv. Neural Inform. Process. Syst.*, 2015, pp. 2017–2025.

[65] J. Tang, J. Wang, and J.-F. Hu, "Predicting human poses via recurrent attention network," *Visual Intelligence*, vol. 1, no. 1, pp. 1–9, 2023.

[66] X.-S. Wei, Y.-Y. Xu, C.-L. Zhang, G.-S. Xia, and Y.-X. Peng, "CAT: A coarse-to-fine attention tree for semantic change detection," *Visual Intelligence*, vol. 1, no. 1, p. 3, 2023.

[67] G. Sun, Y. Liu, T. Probst, D. P. Paudel, N. Popovic, and L. V. Gool, "Rethinking global context in crowd counting," *Machine Intelligence Research*, 2023.

[68] Y. Qiu, Y. Liu, L. Zhang, H. Lu, and J. Xu, "Boosting salient object detection with transformer-based asymmetric bilateral U-Net," *IEEE Trans. Circ. Syst. Video Technol.*, 2023.

[69] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *Int. Conf. Learn. Represent.*, 2021.

[70] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.

[71] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Adv. Neural Inform. Process. Syst.*, 2019, pp. 8024–8035.

[72] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu *et al.*, "MMDetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019.

[73] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[74] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Int. Conf. Comput. Vis.*, 2019, pp. 9627–9636.

[75] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Adv. Neural Inform. Process. Syst.*, 2015, pp. 91–99.

[76] M.-M. Cheng, Y. Liu, W.-Y. Lin, Z. Zhang, P. L. Rosin, and P. H. Torr, "BING: Binarized normed gradients for objectness estimation at 300fps," *Computational Visual Media*, vol. 5, no. 1, pp. 3–20, 2019.

[77] K. Simonyan and A. Zisserman, "Very deep convolutional net-

works for large-scale image recognition," in *Int. Conf. Learn. Represent.*, 2015.

[78] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[79] Y. Liu, Y.-H. Wu, P. Wen, Y. Shi, Y. Qiu, and M.-M. Cheng, "Leveraging instance-, image-and dataset-level information for weakly supervised instance segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1415–1428, 2022.

[80] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.